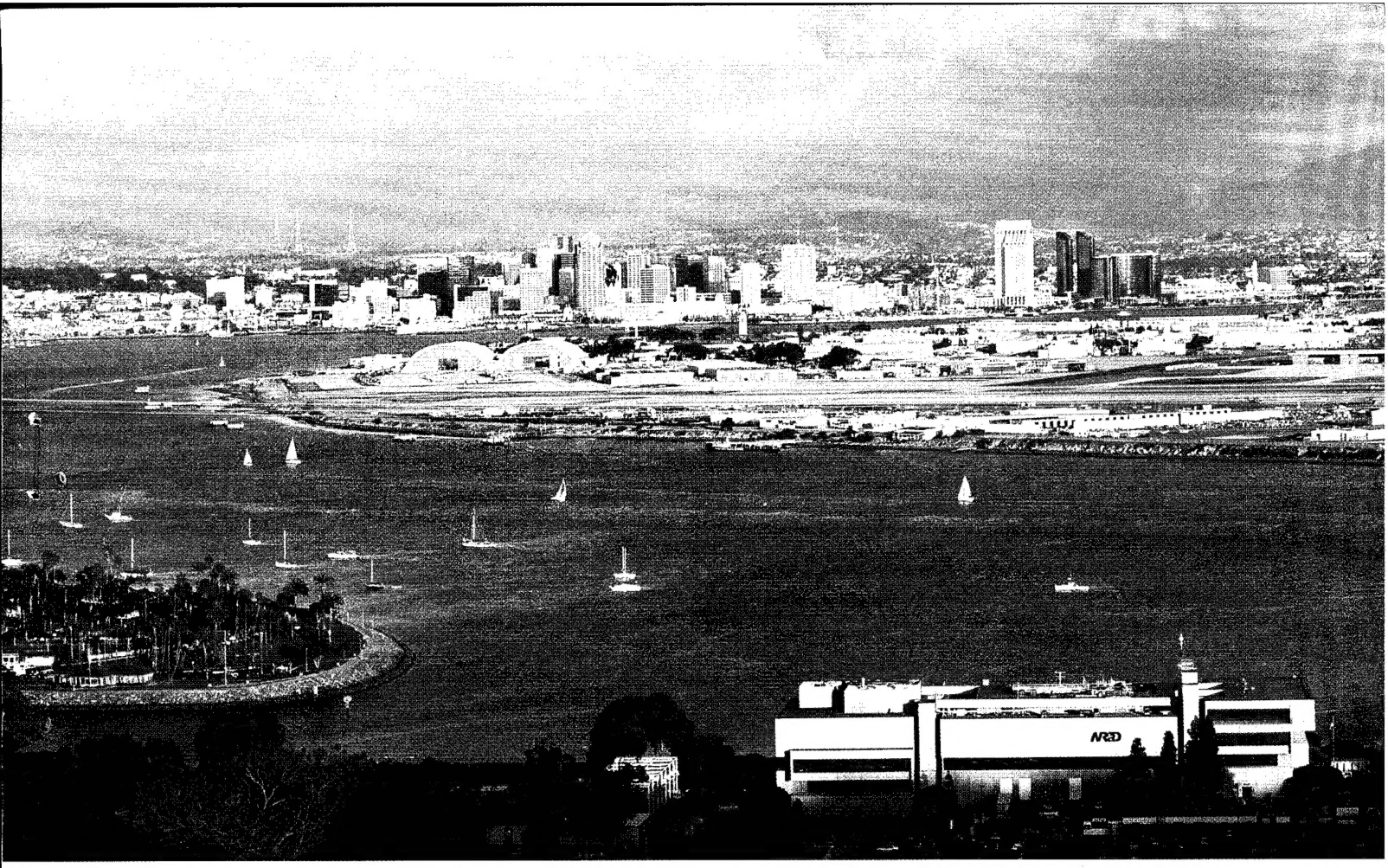


Technical Report 1652
August 1994

Detection of Stress by Voice: Analysis of the Glottal Pulse

Jeff Waters Steve Nunn Brenda Gillcrist Eric VonColln

Naval Command, Control and
Ocean Surveillance Center
RDT&E Division

San Diego, CA
92152-5001

Technical Report 1652
August 1994

Detection of Stress by Voice: Analysis of the Glottal Pulse

Jeff Waters
Steve Nunn
Brenda Gillcrist
Eric VonColln

NAVAL COMMAND, CONTROL AND OCEAN SURVEILLANCE CENTER
RDT&E DIVISION
San Diego, California 92152–5001

K. E. EVANS, CAPT, USN
Commanding Officer

R. T. SHEARER
Executive Director

ADMINISTRATIVE INFORMATION

The work detailed in this report was performed by the Naval Command, Control and Ocean Surveillance Center, RDT&E Division, Speech Technologies Branch, Code 44213, for the Office of Naval Research, Arlington, VA 22217–5000. Funding was provided under project ZF2301, subproject RV36I21, program element 0602936N, and work unit DN 303026. The study was conducted from October 1992 to September 1993.

Released by
S. W. Nunn, Head
Speech Technologies Branch

Under authority of
J. D. Grossman, Head
Simulation and Human
Systems Technology Division

ACKNOWLEDGMENTS

We would like to thank the Office of Naval Research and Dr. Kenneth Campbell, Head of NRaD's Science and Technology Programs Office, for sponsoring this Independent Exploratory Development study. We also want to thank William Nugent and Vladimir Goncharoff, who provided advice and assistance during this investigation.

RBT

EXECUTIVE SUMMARY

OBJECTIVE

The objective of this Independent Exploratory Development (IED) study was to determine whether or not significant measures are present in the human voice for detecting the emotional reaction, "stress." A technique was implemented for automatically measuring parameters of the glottal pulse to see which might be indicators of stress. The glottal pulse is the pulse of air generated at the vocal cords when they open and close. Parameters measured included opening slope, closing slope, pitch, and parameters derived from modeling the pulse with a beta function. The measures were analyzed for significance both across and within speakers.

RESULTS

The results of this IED study confirm that several of the measures are significant indicators of stress; for example, the glottal pulse generally narrows or shifts in mass under stress. And although pitch and beta-function parameters were significant across speakers, the measures were largely speaker dependent.

RECOMMENDATIONS

The measurement technique and analysis could be used for stress detection in appropriate applications where speech from subjects can be collected directly from a microphone. Further development of these techniques in the frequency domain and with higher frequency components of the voice is possible and would be an appropriate follow-on to this effort.

CONTENTS

CHAPTER 1, INTRODUCTION 1

CHAPTER 2, BACKGROUND 3

COMPLEXITY: THE HUMAN VOCAL TRACT 3

SIMPLICITY: THE SPEECH PRODUCTION MODEL 4

CHAPTER 3, STUDY DESIGN 7

CHAPTER 4, QUANTITATIVE MEASURES 9

MEASURE #1: AMPLITUDE 9

MEASURE #2: OPENING SLOPE 10

MEASURE #3: CLOSING SLOPE 10

MEASURE #4: RATIO 10

MEASURE #5: PITCH 11

MEASURE #6: AA 12

MEASURE #7: BB 13

MEASURE #8: CC 13

CHAPTER 5, INVERSE FILTERING 15

CHAPTER 6, EXTRACTION OF DATA: TOOLS AND TECHNIQUES 17

CHAPTER 7, ANALYSIS OF DATA 19

REGRESSION 19

ANOVA 20

MANOVA 21

ANALYSIS PART I: ACROSS SPEAKERS 21

ANALYSIS PART II: WITHIN SPEAKERS 22

CHAPTER 8, RESULTS 25

ACROSS SPEAKERS 25

1. Amplitude—Across Speakers 26

2. Opening Slope 29

3. Closing Slope 31

4. Ratio of Opening Slope to Closing Slope 34

5. Pitch 37
6. Beta Function Measure: AA 39
7. Beta-Function Measure: BB 42
8. Beta-Function Parameter: CC 45
WITHIN SPEAKERS 47
CHAPTER 9, NEURAL NETWORK EXPLORATION 59
CHAPTER 10, CONCLUSIONS 63
CHAPTER 11, REFERENCES 65
CHAPTER 12, BIBLIOGRAPHY 67

FIGURES

1. Speech production model 16
2. IAIF block diagram 16
3. Example glottal pulse 17
4. Possible normal/stress curve paths 20
5. Physical structure of neuron 59
6. Neural network processing element 60
7. Neural network structure used for glottal pulse classification 60

CHAPTER 1
INTRODUCTION

Technology advances, sometimes steadily, sometimes rapidly, and occasionally in giant leaps. The giant leaps may start as merely a small technical advance, but they may occur in an area having major human impact; for example, in the human-computer interface. If this is so, then we are on the verge of a giant leap forward in human interfaces: computer awareness.

Whether or not we believe the computer can think, the computer is on the verge of understanding. Currently, a computer is a machine of metal and electronics, essentially unaware of our existence. The computer will take our input; respond, if we enter commands; and perhaps perform actions on its own, based on a timer. But it neither knows nor cares if the operator is in a bad mood, a good mood, happy, sad, distressed, anxious, worried, or somber. It is unaware of the operator's condition. For this reason, the current computer, despite its awesome capabilities, is but a tool, like a screwdriver or a hammer. It is not a friend, a companion, nor a colleague.

The major leap will come when the computer can respond without being asked. Not responses based simply on a timer or a calendar, but rather responses based on its awareness of the operator's condition. This computer will know when the person is tired and ask if the operator wishes to continue. It will know when the person is stressed and suggest ways of easing the current workload and perhaps assist in prioritizing tasks. It will know when the operator is angry and provide additional help.

How will the computer be "aware" of one's condition? In the same way a person's friends or colleagues are aware, by one's actions: predominantly through the person's body language, facial expressions, and voice. The computer will have a built-in video capability to monitor one's expressions and body language, and it will have voice-recognition and analysis algorithms for understanding and interpreting human voice.

How long until the computer becomes "aware" of a person's condition? Currently, computers are good at number crunching, but not at human analysis; for example, face recognition, voice recognition, and the ability to focus on the important details. Fuzzy-logic, neural networks, and parallel processing are all techniques for enabling a computer to think more like a human. Still, these techniques are under development, and 20 or more years may pass until a reasonably aware computer is commonly available. But, fortunately, the small steps in this large area are already underway and are appearing in interfaces. These steps include limited voice recognition, built-in voice recording, and even a built-in video capability. The link currently missing is the software and associated algorithms for analyzing video images and voice cues to a person's emotional state.

This report discusses an attempt made to discern whether or not significant measures are present in the human voice to detect one's emotional reaction, stress. Although these efforts investigated only one emotional state, the technique was based on research that suggests the technique could be applied to detect other emotional states. The current study was focused on trying to find significant measures, but a secondary concern was to try to derive these measures by an automatic means that could be implemented by a computer. Although this is only one step, it suggests implementation of a technique a computer can use to automatically identify stress. The impact of such a technique on human operators could be immense.

CHAPTER 2
BACKGROUND

Speech is a complicated, subtle, and amazing activity. Furthermore, speech and hearing are intertwined in important ways, so analyzing one without the other is difficult. Nonetheless, for this report, speech is discussed from the standpoint of production, that is, how speech is produced—not how it is heard. First, some of the complexities of speech production are described, followed immediately by a "model" of the speech-production process. This model captures essential elements of the process and ignores many of the details. The model has proven its usefulness in real-life applications, suggesting its validity in the way it represents the essential elements. However, the model is not perfect; therefore, problems and nonideal results can be expected in its implementation.

The following discussion covers the mechanics of speech production and intends only to suggest the complexity of the speech-production process. This background information will be helpful in understanding the subsequent description of the speech-production model.

COMPLEXITY: THE HUMAN VOCAL TRACT

Speech is generated by the interaction of several component pieces of the vocal tract. The following discussion summarizes this process in a simplified manner.

The vocal cords act essentially as a valve that opens and closes in the lower part of one's throat. This allows air to be pushed through the *glottis,* which is the opening between the vocal cords. In a normal, relaxed state, the glottis is open. This is the normal state when breathing, rather than speaking. When the vocal cords vibrate, the glottis opens and closes rapidly. The rate of vibration is the fundamental frequency of one's voice, which listeners detect as the "pitch" of that voice.

Our vocal cords don't always vibrate when we are talking, they only vibrate for voiced sounds, like vowels. During unvoiced sounds, the glottis remains open.

The "glottal pulse" is the wave that corresponds to a single pitch period, that is, a single opening and closing of the glottis. The specific manner in which the glottis opens and closes imparts a specific shape on this pulse. Each glottal pulse is like a puff of air. A series of these glottal pulses are generated for any given voiced segment, and the entire series is referred to as the "excitation" signal. This name comes from engineering terminology, where a signal excites a filter, as described in the following discussion.

This excitation signal (series of glottal pulses) moves up through the throat and eventually exits out the mouth and nose. Somewhat surprising is that air passing through the nose is a major contributor to speech. However, this is quite commonly noted, for example, when we say a person has a "nasal" voice. Here we are detecting in the sound of their voice a dominant resonance in the nasal cavity.

As the excitation signal moves upward on its way though the mouth and nose, it encounters obstructions. First, the walls of the throat (surrounding the *pharyngeal cavity*) impede its progress. This impedance causes certain resonant frequencies in the signal as it bounces off the walls of the throat. This is similar to a person yelling in a cave; certain echoes and resonances occur. The same effect is caused by the walls of the mouth surrounding the *oral cavity* and by the walls

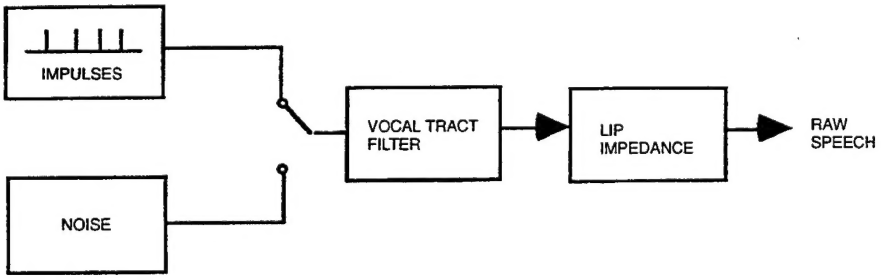
of the nose surrounding the *nasal cavity.* The sizes and shapes of these cavities vary for each person, and the resonances generated are in some ways, speaker specific. The resonances of these three cavities (pharyngeal, oral, and nasal) are often referred to as formants—the first formant, the second formant, and the third formant respectively. Higher formants are possible, because the vocal tract is more complicated in its shape than simply three separate definable cavities, and because of the movement of articulators, described next.

In addition to the impedance caused by the walls of the cavities in the vocal tract, the excitation signal runs into other obstructions as it moves through the mouth and nose. These obstructions include parts of the vocal tract that one controls, including the tongue, velum, jaw, and lips. The velum is a piece of skin in the upper back of the throat, which acts essentially as a valve into the nasal cavity. If the velum is open, air moves into the nasal cavity. If the velum is closed, no air moves into the nasal cavity; instead, it all comes out of the mouth. When one is at rest, breathing through the nose, the velum is open. The air passes through the nasal cavity, made up of many passages lined with mucous tissue, which severely dampens any sound passing through.

These additional obstructions (tongue, lips, teeth, velum, and jaw) are often referred to as "articulators," because we move them in certain ways when articulating or speaking. The tongue, for example, is an amazing instrument for speech, with muscles controlling its height and shape independently for the back, middle, and tip. The lips control the size and shape of the mouth opening. They can be extended or withdrawn, widened or narrowed, opened or closed. The tongue and the lips are the major articulators, because they have the greatest influence on the shape of the mouth and the type of sound generated. The jaw aids in positioning the tongue and lips, and the teeth are used to create friction with the "m." The alveolar ridge, on the roof of the mouth just behind the front teeth, is often used as a major contact point for the tip of the tongue for many sounds. The velum is closed for most speech sounds, allowing no air to move through the nasal cavity; however, it opens for nasal sounds, such as "m" or "n."

SIMPLICITY: THE SPEECH PRODUCTION MODEL

Some 20 years ago, engineers studied the complexity of the vocal tract and tried to model it. Their idea was simple, but useful. The vocal tract can be modeled by three components: (1) an excitation signal, (2) a vocal tract filter, and (3) a lip radiation. The engineers used engineering terminology (signals, filters, radiation), but provided simple drawings to explain the concepts.



The excitation signal takes one of two forms. If the vocal cords are vibrating (as with voiced sounds, like vowels), a periodic fundamental frequency (pitch) is being generated. In this case, the excitation signal can be modeled as a series of periodic (equally spaced) impulses or spikes, as if one were beating a drum. If the glottis is open (as with unvoiced sounds, like "s" or "sh"), then there is no periodicity, and the air passes randomly by the glottis. In this case, the excitation signal resembles noise; it has no characteristic tone.

The vocal tract filter is simply the result of imagining that the vocal tract is a series of tubes of different sizes. One tube might represent the pharyngeal cavity, while another represents the nasal cavity, and so on. They are different sizes because of the different effects caused by different parts of the vocal tract. The engineers developed methods for estimating the areas of these different tubes and for deriving numbers (coefficients). These serve as weights to determine the effect of each tube on the output speech at any given moment in time. For example, linear predictive coding (LPC) is one method for estimating the shape of these vocal tract tubes. The LPC coefficients can then be used as a "filter."

A "filter" is essentially a series of numbers, perhaps 14, which, when multiplied in a special way with a signal (convolved), can change the signal (air-wave shape) into another signal. Once LPC coefficients are derived from a speech signal, an excitation signal (for example, a series of periodic pulses) can be sent into the LPC filter. The filter will then change the excitation signal into a speech signal resembling the original speech for that moment in time. This is quite an achievement, and LPC is now commonly used in many applications, including the STU-III secure communication phones used in military installations.

Another application of the LPC filter is evident if the filter is turned upside down, that is, if the filter is inverted. Before, with the filter rightside-up, an excitation signal could be converted into a speech signal. Now, by inverting the filter, the opposite can be done; a speech signal can be converted into an excitation signal. This is exactly what is wanted in this study. The following coverage will show that the focus of this study is on the excitation signal. The excitation signal needs to be derived from the operator's speech by inverse filtering. This will be covered in more detail later.

So far, this discussion has covered two of the three components of the speech production model, namely, the excitation signal and the vocal tract filter. The third and final component is an adjustment to account for lip radiation. After the excitation signal passes through the vocal tract, it must leave the lips (and the nose). The lips impede the signal, and the more they are closed, the more they impede the signal. The radiation from the lips has the general effect of boosting the higher frequencies in the signal. This effect needs to be accounted for in the speech production model because it significantly affects the resulting sound heard by the listener. Engineers have modeled this effect by a differentiator and, consequently, the inverse effect is an integrator. Of particular interest in this study is the inverse effect of lip radiation and the integrator, because to extract the excitation signal from the speech, two steps must be inverted: (1) inverse filter the speech with the vocal tract filter and (2) undo the effect of lip radiation by applying an integration.

In summary, the speech production model assumes that a speech signal can be generated with three components, the excitation signal, the vocal tract filter, and the lip radiation. For reasons discussed below, this study focuses on the excitation signal. The goal here is to extract the excitation signal from the operator's speech. To do this, the speech is assumed to be produced by a system that follows the speech production model. Therefore, to extract the excitation signal, the effects of the vocal tract filter and the lip radiation must be removed from the speech. An attempt is made to eliminate the vocal tract effects by estimating the vocal tract filter and then inverse filtering the speech to remove the effects of the vocal tract. Finally, the signal is integrated to remove the effects of the lip radiation. When finished, this provides an estimate of the excitation signal; not a perfect estimate, but hopefully good enough to allow the effects of stress to be measured.

CHAPTER 3
STUDY DESIGN

Ideally, we would have liked to conduct experiments of our own design to collect speech under normal conditions and conditions simulating typical Navy command and control stress. In such an experiment, physiological measures of stress would have been used, including heart rate, galvanic skin response, blood pressure, and similar physical measures to confirm that stress was induced. Although physiological measures are not always correlated with stress, the first attempt at finding a stress measure in the voice could well have been based on the following concept: a stress measure should be looked for first with stress that had corresponding physiological correlates. Later, we could look for measures in the voice in stressful conditions, regardless of correlates.

More could be said concerning this; and in fact, a followup effort is recommended in this regard. For now, we can simply note that the scope of the current study did not allow for such experimental collection of data. Instead, we extracted our data from a previously collected database of speech sounds.

The Lincoln Laboratories Speech Style Database was used in this study. This database consists of 9 speakers, each of whom spoke 35 words, twice, under various speech styles, including angry, loud, soft, and other conditions. Two of these conditions were the subject of this study: (1) Normal and (2) 70-percent task-loaded (referred to as "cond70"). The 70-percent task-loaded speech was collected while the speakers were engaged in an eye-hand coordination task, similar to a task requiring manipulation of a joystick while tracking a dot on the computer screen. Some physiological measures were collected as well, but did not appear to correlate strongly with the stressful condition. The physiological measures were not used in this study.

The database included speech from 9 speakers, each speaker saying 35 different words twice under each condition. We analyzed one glottal pulse from each utterance. Ideally, we could have analyzed 70 glottal pulses for each speaker for each of the 2 conditions, normal and cond70. (Thirty-five words per speaker $\times$ 2 utterances per word $\times$ 1 glottal pulse per word = 70 glottal pulses per speaker for each condition.) In fact, our study used less than half of these pulses (29) for some parts of the analysis, due to limitations in implementing the "automatic" extraction and measuring process.

Two problems arose during automated extraction, which limited the size of the data set. First, the inverse filtering, which extracted the excitation signal from the speech, was sensitive to the total size of the voiced segment selected. If the voiced segment was too short, the inverse-filtering algorithm failed. This was not an insurmountable problem, but time constraints required manual intervention to select segment size, rather than fine tune the automated algorithm. Second, the automatic pitch-detection algorithm failed on several occasions, which is not surprising, since pitch detection is a difficult task.

In the end, data were extracted from at least 29 pulses for each speaker, under each condition, for a total of 29 pulses $\times$ 9 speakers $\times$ 2 conditions = 522 pulses analyzed. This amount of data proved sufficient for our purposes. That is, the within speaker analysis did not require all speakers to have the same amount of data, since we were not comparing across speakers. For this part of the analysis, the lowest number of pulses analyzed per speaker per condition was 29, and more were analyzed if available. The largest number available for any one speaker was 48 per condition.

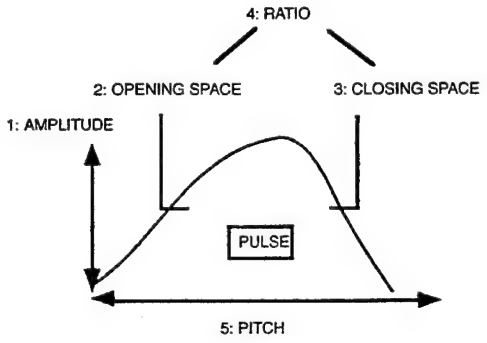
In the data file, each row represents one entry, i.e., one set of measures for one glottal pulse.

Analysis by a computerized statistical package could compare the mean values of each measure. This was done across speakers and within speakers for the two conditions to see if a significant difference existed. A later section of this report details the statistical analysis approach. Also, the next section of the report covers the measures (amplitude, opening slope, etc.), and how they were chosen and measured.

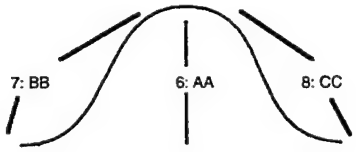
CHAPTER 4
QUANTITATIVE MEASURES

Voice has been the subject of research efforts concerning stress for several years. A listing of several of these papers is provided in Chapter 12, Bibliography. This study, however, was largely inspired by work performed at Georgia Tech on the excitation signal (Cummings & Clements 1990, 1992). In that work, researchers at Georgia Tech concluded that the styles of speech in the Lincoln Laboratory Speech Style Database were evidenced by characteristic changes in the shape of the glottal pulse. The studies were somewhat limited, however, both in terms of the number of speakers evaluated and the complexity of the algorithms used.

Recent research by Kathleen Cummings and Mark Clements at Georgia Institute of Technology analyzed glottal waveforms of various speech styles. Their conclusions suggested that styles of speech could be distinguished by measuring various parameters of the glottal pulse, as well as by looking at parameters generated by a beta-function model of the pulse. Based on these findings, this study attempted to apply a set of these parameters to focus on identifying stressed versus normal speech. The parameters are shown in the diagram and their characteristics are described below:



EIGHT GLOTTAL PULSE PARAMETERS



MEASURE #1: AMPLITUDE

The loudness of a speaker's voice has intuitive appeal as a measure of stress. It is logical to imagine that a speaker's voice might get louder when the speaker is under stress. Similarly, a different speaker might respond with a quieter voice under stress. Either way, a significant shift in loudness might be an important characteristic of a person's emotional state.

In addition to its intuitive appeal, another important reason exists for measuring amplitude. The amplitude of a pulse affects other parameters of the pulse, such as opening and closing slope. Two identical pulses, except the first louder than the second, will show different slopes.

The louder pulse will have steeper slopes than the second pulse, because it is taller, that is, higher in amplitude.

This discussion will demonstrate that opening and closing slope are potentially significant parameters. In order to measure them correctly, without having them polluted by the effect of different amplitudes, the pulse amplitudes will have to be normalized. In other words, the pulses must be adjusted so they all have the same amplitude. To perform this normalization, the amplitude must be measured, and it can be analyzed at the same time.

Amplitude measurement can be performed by simply measuring the height of the pulse, with normalization then performed by dividing every point in the pulse by the maximum value in the pulse. This results in a normalized pulse with highest point equal to 1.0.

MEASURE #2: OPENING SLOPE

Glottal pulses are created by air pressure released by the opening and closing of the vocal cords. The speed of the opening and closing can vary, with faster opening or closing correlating with steeper opening or closing slopes of the glottal pulse. In normal speech, the opening slope is usually more gradual than the closing slope.

Stress effects may cause tension in the vocal cords, which could correlate with faster opening or closing slopes. For this reason, these characteristics should be measured.

Both amplitude and pitch affect the opening and closing slope. A louder pulse will have steeper slopes, as will a pulse of significantly higher pitch. However, amplitude and pitch can be measured separately, so removing these effects on the slopes is preferable. Normalizing the pulses for pitch and amplitude would allow us to measure slope changes that relate directly to pulse shape changes caused by other effects, such as stress. For this reason, both amplitude and pitch are normalized before measuring slopes. Amplitude normalization was discussed above; and pitch normalization is discussed below, during the discussion of the pitch measure.

One question is where to measure opening slope, which varies from the bottom left start of the pulse to the top. Then, the question arises: where on the rising edge of the pulse should the slope be measured? We decided the best measurement spot would be the halfway point, that is, halfway up the rising edge of the pulse. This spot can be determined by selecting the midpoint between the first point of the pulse and the point of the pulse corresponding with the top of the pulse. At this location, a specific slope measurement was taken by simply measuring the slope on one point on either side of the midpoint.

MEASURE #3: CLOSING SLOPE

Similarly, closing slope is of interest, due to its correlation with the rate of closure of the vocal cords. Any tension or related stress effects might well affect this slope.

As with opening slope, both amplitude and pitch are normalized prior to measuring the closing slope. The closing portion of the pulse is measured at the midpoint, halfway between the pulse peak and the last point in the pulse.

MEASURE #4: RATIO

Another way in which the slopes may vary is in relationship to each other. For example, if a stress effect were to cause a disproportional impact on closing slope over rising slope, a change in the ratio would be expected.

In this case, the ratio of opening slope to closing slope is computed. These two slopes have been measured, so the computation is a simple process.

MEASURE #5: PITCH

Pitch is the rate of vibration, i.e., the opening and closing of the vocal cords. In male speakers, an average pitch is approximately 120 Hz, and for female speakers, 220 Hz. Pitch is an easily noticed effect, providing the fundamental tone of a person's voice. The sensitivity of listeners to pitch effects and the direct correlation of pitch with the creation of glottal pulses suggest this would be a fruitful measure for stress effects.

Pitch was measured by inverse filtering a number of pitch periods of the input signal, then performing a cepstral analysis on the resulting glottal-pulse train. The pitch was estimated from the cepstrum.

The cepstrum is a standard representation of a speech signal, computed by performing a discrete Fourier transform on the log of the signal magnitude. Several texts discuss this technique and its application for pitch detection.

After measuring the pitch, we attempted to normalize the pulse train to a standard pitch of 140 Hz, with the goal of eliminating any slope effects caused by pitch. Normalization was achieved by interpolation and/or decimation on the glottal pulse train.

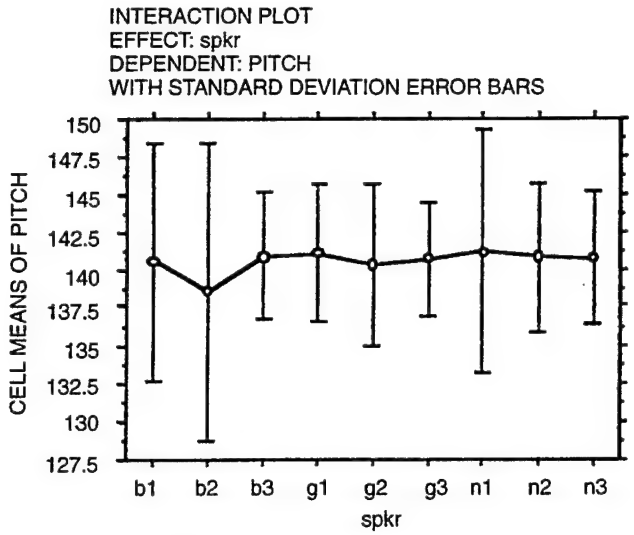
Although normalization was not perfect, automatic pitch detection was applied on a broad set of normalized pulses to test the effectiveness of the normalization. Mean values of both stressed and normal normalized pulses were found to be 140, with a standard deviation of approximately 9 Hz. Any error in normalization appeared to be spread equally among the stressed and normal pulses. Although this was not a formal test, it may provide some evidence of the performance of the pitch normalization. We performed an ANOVA using pitch as the dependent variable; and speaker, condition (stressed or normal), and the interaction of speaker and condition as independent variables. If the normalization were ineffective, a significant variation in mean pitch would be expected for speaker or condition. None was found, as evidenced in the ANOVA table below:

Type III Sums of Squares

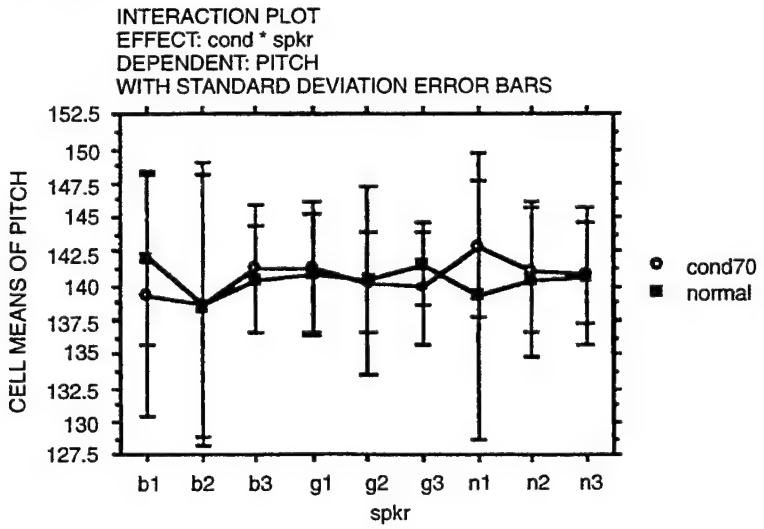
Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	366.455	45.807	1.160	.3210
cond	1	3.139	3.139	0.079	0.7781
cond * spkr	8	499.298	62.412	1.580	.1269
Residual	722	28514.278	39.493		

Dependent: pitch

The lack of significance suggests that the normalization was at least reasonably effective. The normalization results can be further reviewed by plotting the mean pitch measured on the normalized pulses by speaker. If the normalization were ineffective, a plot with some significant variation would be expected. Here is the plot of mean pitch by speaker for the normalized pulses:



The plot is relatively flat, with little variation, and all means centered within 3 Hz of 140. This suggests a reasonable normalization performance. The normalization can be further reviewed by plotting the pitch means by condition (cond70 or normal), as well as by speaker. If normalization is effective, no significant variation would be expected between the mean values of stressed and normal speech for each speaker. Here is the plot of mean pitch by condition and speaker for the test set of normalized pulses:



The plot does show some variation between means for stressed and normal pulses for two or three speakers; but as the earlier ANOVA confirms, the variation is not significant. The plots provide some indication that the pitch normalization is reasonably effective.

MEASURE #6: AA

The measure, AA, along with the next two measures, BB, and CC, is a parameter derived from a beta-function model of each pulse. A beta function is defined as:

$$y = AA * (x/xmax)^{BB} * (1 - (x/xmax))^{CC}$$

The parameter, AA, controls the amplitude or height of the beta-function pulse. BB affects the slope of the rising portion of the beta-function model, while CC affects the slope of the closing portion of the model.

Every glottal pulse was modeled with the beta function by performing an iterative "best fit," that is, least mean square algorithm. The beta function is a smooth function, so it does not model local small changes in parts of the rising or closing slope; instead, it models the overall pulse shape. It is sensitive to shifts in the pulse mass and other global shape changes, such as pulse narrowing or widening.

All the glottal pulses are normalized for both amplitude and pitch prior to finding the "best fit" beta-function approximation. In fact, an additional pitch normalization was applied only for these three parameters, AA, BB, and CC. This additional normalization was applied individually to each modeled glottal pulse. Interpolation and/or decimation was performed so that all pulses were normalized to exactly the same length. This was done to further eliminate any glottal-pulse shape effects due to pitch.

In the beta-function model, AA controls the height of the pulse model. Yet AA is not independent of BB and CC. For example, if BB and CC are higher for one pulse, then AA must also be higher to maintain the normalized amplitude of the original pulse. In other words, AA is affected by BB and CC, and in some ways is a more sensitive measure of BB and CC.

MEASURE #7: BB

The measure, BB, controls the slope of the rising portion of the beta-function model of the glottal pulse. Any significant change in rising slope should be reflected by this parameter, assuming the change is a major change, not just a local ripple in the rising portion of the pulse.

The hope is that beta-function parameters, such as BB, can capture overall changes in the glottal pulse due to stress. Otherwise, this might be missed by local, specific measures, such as our midpoint measures of opening and closing slope.

MEASURE #8: CC

CC controls the slope of the closing portion of the beta-function model of the glottal pulse. If the vocal cords shut more quickly or slowly under stress, an overall change in the slope of the closing portion of the waveform would be expected. CC may measure this affect.

In summary, some general observations can be made. First, all measures are in the time domain. This means that maintaining accurate phase relationships of the original signal frequencies is important when collecting data. Random phase changes induced by a typical communication channel would make these types of measures difficult to gather. Direct data collection is suggested for any system relying on these measures. Second, frequency-domain measures may be equally reliable; however, the focus of this study was on the time-domain representation. This is due to previous successful analysis of time-domain features for style detection by other researchers. Further research effort is recommended for comparing frequency characteristics of the glottal pulse.

A third observation is that normalizing the pulses for amplitude and pitch became a significant portion of the effort, due to the desire to collect "pure" glottal-pulse measures. In other words, we wanted to see if the vocal cords are affected by stress in some way that changes the shape of the pulse, regardless of how loud someone talks or of the basic tone of their voice. The goal was to attempt to get the essence of the pulse.

The next discussion (Chapter 5) covers our basic technique (Inverse Filtering) for extracting glottal pulses from speech; it then describes our tools and techniques for automatic processing. The sections that follow cover the measures for data collection in more detail. And, in addition, these measures are the subject of significant statistical analysis discussed in Chapter 8.0, Results.

CHAPTER 5
INVERSE FILTERING

The method of glottal-pulse extraction used is based on the Iterative Adaptive Inverse Filtering (IAIF) method (Alku, 1992). This method is chosen because it is automatic, that is, no operator is needed to tell it where to find a 'good' glottal pulse; additionallly, the algorithm is easily implemented.

The IAIF method assumes a basic speech model, as in figure 1. Here, the idea is that a train of impulses is used to drive the filter for voiced sounds, like vowels; and the white noise is used to drive the filter for unvoiced speech, like fricatives. The input to the vocal tract filter is switched between these two, depending on whether the speaker is producing voiced or unvoiced speech. The pulse trains are analogous to the glottis opening and closing (nonideal), and the white noise generator is analogous to the situation where the glottis is held open instead of opening and closing. The filter is analogous to the effects the driving function incurs moving through the throat and into the mouth (vocal tract). The last block represents the impedance caused by the lips and is usually modeled as a single or double differentiation. Although the glottis waveform is not an impulse (it has a rise and fall time), the IAIF method assumes this model and works backward from the raw speech to try to obtain an estimate of the glottal pulse.

A diagram of the IAIF method can be seen in figure 2. In block 1, a high-energy component of the input speech is found and filtered with a high-pass filter to eliminate any dc bias that might be present, that is, accumulated at the integrator. A high-energy part of the speech is used because that is when voiced speech occurs and, therefore, it is when the glottal pulse should be estimated. In block 2, an LPC fit of order 1 is used to estimate the contribution of the glottal pulse to the speech. The idea here is that the glottal pulse is a slowly varying component (that is, low frequency), so a 1-pole LPC fit will tend to track this component. Once the pulse is estimated, the LPC coefficient is used to inverse filter (block 3) the speech in order to remove the effects of the glottal pulse. The input to block 4 is now a representation of the raw speech, with the glottal pulse component filtered out. Block 4 now tries to estimate this speech in fine detail (high-frequency component) with a 12-pole LPC model. In block 5, these LPC coefficients are then used to filter out the effects of the vocal tract, leaving an estimate of the derivative of the glottal pulse.

To obtain an estimate of the glottal pulse, integration would be performed twice to account for the impedance of the lips. But to obtain a better estimate, the process is repeated (in blocks 6–10), starting out with, presumably, a better measure of the glottal contribution. In block 6, a second estimate of the glottal pulse is taken with a 4-pole LPC model. These coefficients are used to inverse filter the original speech. The input to block 8 is now raw speech, with the glottal contribution filtered out. This is then modeled with a 12-pole LPC fit, and again the raw speech is inverse filtered, leaving an estimate of the derivative of the glottal pulse. Block 10 then integrates the signal twice to account for the impedance of the lips. The output of block 10 is an estimate of the glottal pulse.

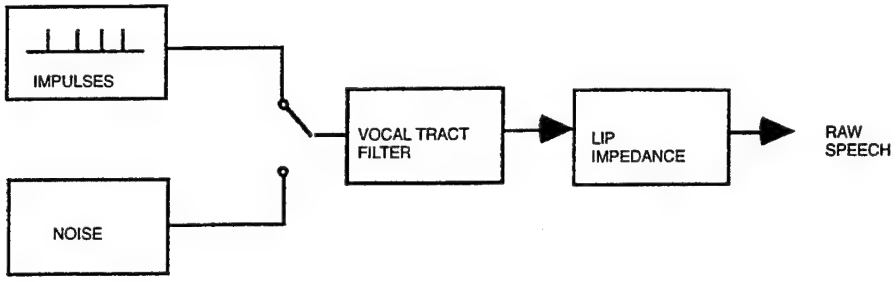


Figure 1. Speech production model.

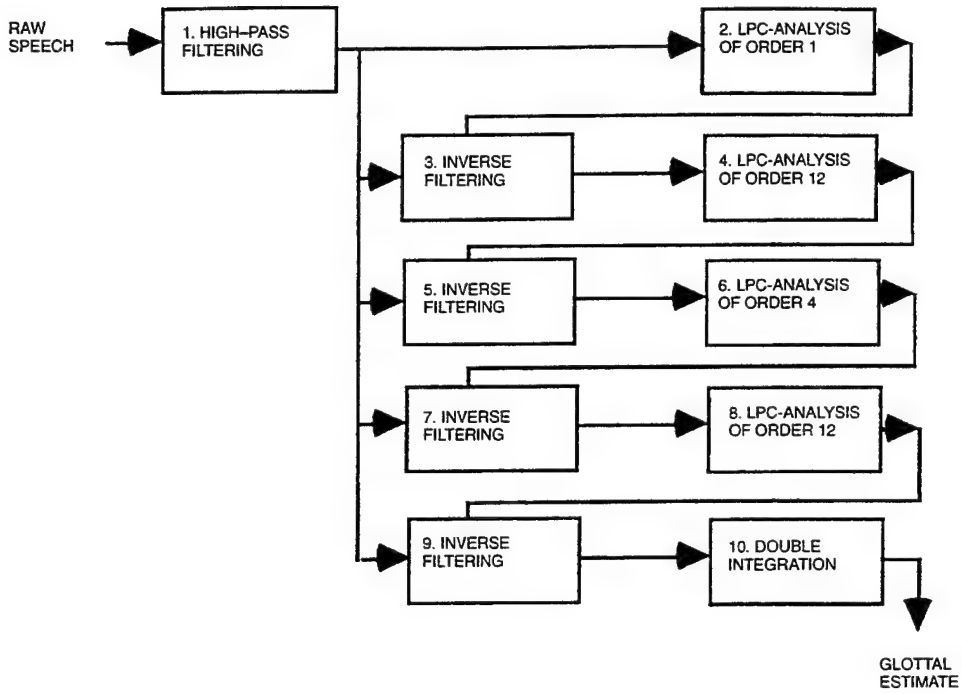


Figure 2. IAIF block diagram.

CHAPTER 6
EXTRACTION OF DATA: TOOLS AND TECHNIQUES

In figure 3, a time waveform represents the excitation signal at the glottis for one speaker; and one word and one condition is the input to the program that extracts a single glottal pulse. Computed here are the autocorrelation of the time waveform and the first and second derivative of the autocorrelation. The first derivative is an indicator of the peaks (maximum and minimum points) of the function. The sign of the second derivative indicates whether the peak found by the first derivative is a maximum or a minimum; and the first and second derivatives are used to determine the first peak of the autocorrelation function. The distance to the first peak in the autocorrelation indicates the period. The original glottal-time waveform is then subdivided into smaller arrays, based on the length of the period. The index of the minimum value of each subarray is detected and stored in an array. A single pulse was extracted by picking the median indices in this array and using these indices to pull a subarray out of the original glottal-time waveform. This single pulse is normalized by amplitude, and then these data are used to calculate the feature parameters.

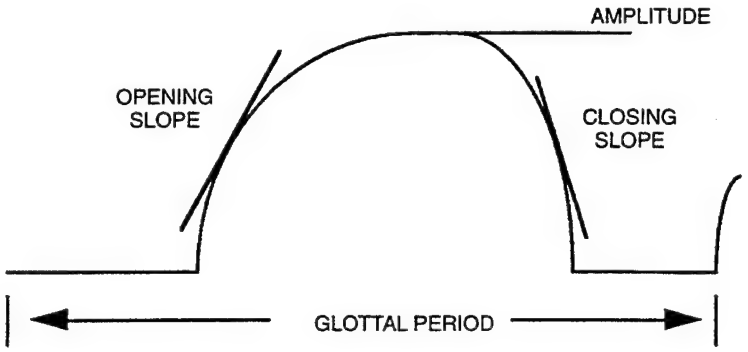


Figure 3. Example glottal pulse.

The parameters measured on the single glottal pulse are the opening slope, closing slope, ratio of the opening slope to closing slope, and the amplitude and the pitch. Amplitude (maximum value) is measured before the pulse is normalized by amplitude. Pitch is measured by finding the peak in the cepstrum of the pulse train within the proper pitch boundaries. The pitch is then normalized out by proper resampling.

The opening slope is calculated by looking at the left half of the single glottal-pulse data. The midpoint of this portion is determined, a value above and below that point is chosen, and the central difference is calculated. The equation that calculates the opening slope is:

$$os = [(midpoint+1) - (midpoint-1)] / (2/fs) \qquad fs=16000$$

If this slope is negative, then the central difference is calculated about the 1/3 point instead of the 1/2 point. The closing slope is calculated in the same manner by looking at the right half of the single glottal-pulse data. The ratio is calculated by the equation:

$$ratio=opening\_slope / closing\_slope.$$

Three other features are included in the feature set that the statistical analysis is based on. These features are the beta model parameters used by Cummins & Clements (1990, 1992). They

found that the characteristics of the beta function enable it to model the different slopes found across the speaker styles (soft, angry, loud, etc.). The beta-function equation is defined as:

$$y = A * (x/xmax)^{**}b * (1-(x/xmax))^{**}c$$

The amount of symmetry in the pulse shape and the relative steepness of the slopes are determined by the values of b and c.

Through interpolation, each individual pulse is adjusted to be the same length. The tilt of the glottal pulse is also nulled, so the pulse starts and ends at zero amplitude. These constraints are necessary, since the beta-function model is constrained to start and end at zero. The objective is to find the A, b, and c parameters in the equation that minimize the least squared error to the given glottal pulse. The Nelder-Meade simplex algorithm is used to search for the minimum of this multivariable function.

A single pulse (median pulse) is extracted from this waveform. A median pulse was chosen for two reasons. First, the pulses in the middle of the waveform were more representative of the familiar pulse shape. Secondly, researchers at Georgia Tech had chosen the median pulse, therefore, we wanted to be consistent with previous research in this area. The feature calculations were measured on the single pulse.

LabVIEW©, a graphical programming language, was used to develop software programs for the single-glottal pulse extraction and for associated glottal-pulse feature calculations. MATLAB© was used for the FFT magnitude feature calculation and for the beta-function modeling. MATLAB was used in the beta-function modeling, since it can minimize the nonlinear function. The LabVIEW and Matlab programs are included in Appendix A.

CHAPTER 7
ANALYSIS OF DATA

Before describing the specific analyses performed, the following discussion covers the statistical analysis techniques considered for this study and why certain techniques were chosen over others. This background information will help to answer some common questions regarding basic statistical techniques.

REGRESSION

Regression was considered for this analysis, but was ultimately rejected as inappropriate. Regression is simply a way to find the best fit line to the data. The equation for the line is of the form:

$$Y = A_1X_1 + A_2X_2 + A_3X_3 + ...$$

where in our case, Y is the dependent variable, the amount of stress, and X_i are the independent variables, the measures, and A_i are the coefficients (how much weight to apply to each measure). Regression analysis finds the coefficients that make the equation the best fit for the given data. If the line has a slope, then that indicates a significant correlation between at least some of the measures and the amount of stress. In addition, the coefficients can be plugged into the equation along with new measures, and it can generate a new Y value that can be used to predict the amount of stress in the given new sample.

Although, in many applications, regression is the correct technique, in our situation, it is not. This is because our independent variable, Y, the amount of stress, is not a gray-scale value. Instead, our data only have two values for Y, 0 or 1; where 0 means no stress (normal) and 1 means stress (cond70). We have no data for anything in between; for example, we have no data for values of 0.2, 0.4, 0.87, etc., which might indicate degrees of stress.

Why does this lack of intermediate stress data make regression inappropriate? The short answer is that because there are no intermediate data, there is no knowledge of how the line curves between the values of 0 (normal) and 1 (cond70). When we start to use the equation for predicting stress, it is obvious that the new input data, a new set of X_i values, will never result in an exact value of 0 or 1, but instead, will result in some value in between. Now, intuitively, one might think that a value of 0.98 can be treated as a 1; and perhaps a value of 0.2 can be treated as a 0, meaning no stress; or perhaps one could conjecture that we have a "little" stress. But we have no justification for making any such conclusion, because there are no intermediate data giving us any information about the shape of the curve between 0 and 1. Why is the curve so important?

Figure 4 consists of three curves that assume a significant difference exists between stressed and normal for the measures. This occurs because the lines have a slope and could all be possible paths the line takes on along the way from 0 to 1. Assume that a value of 0.1 is obtained for a new data entry. If the path follows curve 1, one would conclude that 0.1 means stress exists, but if the path follows curve 2 or 3, then stress does not exist. Similarly, if a value of 0.5 was obtained for Y for a new data entry, the following conditions can occur. If the path follows curve 1, stress exists; if it follows curve 3, there is no stress; and if it follows curve 2, there is medium stress. The problem is, different results are obtained for different paths, and no data exist to justify picking one path over another.

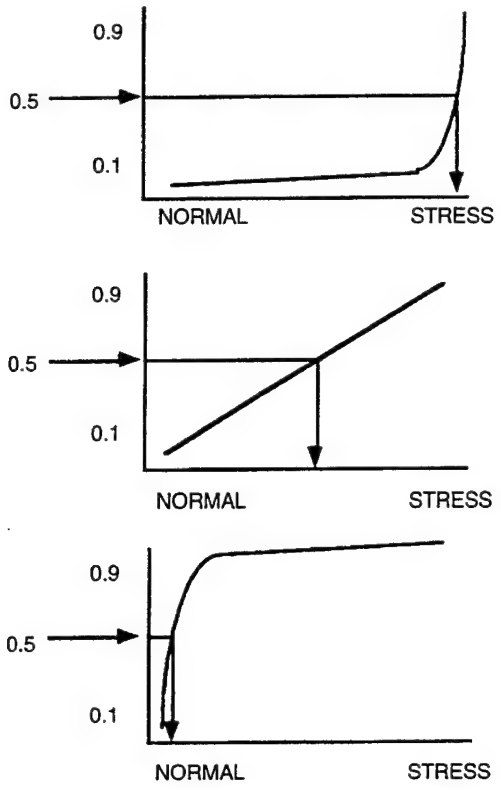


Figure 4. Possible normal/stress curve paths.

Some useful conclusions could be made with regression, but they are limited. One might deduce that significance exists, because a slope exists, but one could not go forward to use the equation for prediction. Conversely, one could perform the regression on a limited data set for training data; and then just assume a linear or some other curve shape and analyze some new data to see how well the prediction worked. If it worked well, one could assume the guessed curve shape was reasonable. However, a strong temptation would exist to adjust the curve shape to fit the data, which is not justified.

Because of the inappropriateness of the regression analysis in our case, we chose to pursue other methods. Even though regression could be used in a limited manner, that information can be gathered through other techniques.

ANOVA

An ANalysis Of VAriance (ANOVA) is a standard technique for determining whether any differences are significant in the mean values of a dependent variable across some independent variable. The null hypothesis being tested is that the mean of the dependent variable is the same, regardless of the level of the main effect. ANOVA models are used instead of regression, when an independent variable has a non-numeric value, such as condition.

A benefit of ANOVA models is their ability to test interactions between factors. An interaction means that the effect of one factor differs depending on the level of another factor. For example, in our case, we could test whether the effect of a condition on a dependent variable, such as pitch, differs depending on who is the speaker. Determining this effect might not be possible simply by looking at a speaker or condition alone.

The problem with an ANOVA, as it pertains to our study, is as follows. If one is performing several different types of measures on the same data sample, as we are when measuring amplitude, pitch, closing slope, etc., on each glottal pulse, then a separate ANOVA must be performed on each measure, i.e., each dependent variable. Yet, the measures may correlate, and this correlation is not being addressed by the separate ANOVA models. In short, your measures may all be measuring the same thing. So when performing multiple different measures on each data sample, you should perform a statistical analysis that considers this correlation. That is, you should perform a Multivariate ANalysis Of VAriance, a MANOVA.

MANOVA

When more than one dependent variable of interest is measured for each data sample, a Multivariate Analysis of Variance is appropriate. For example, in our case, we are measuring opening slope, closing slope, amplitude, etc., which are all variables dependent upon the stress condition and/or speaker. And we are also measuring them on each data sample, that is, each glottal pulse. Note that these measurements are all qualitatively different from each other and that each data sample, each glottal pulse, is measured only once. If the same measure were repeatedly sampled on the same data sample, then a repeated measures analysis might be appropriate.

The details of exactly how a MANOVA is performed will not be covered. Instead, our computer statistical package will be used for the computations. In our case, we are using Super-ANOVA™ to perform the statistical computations.

ANALYSIS PART I: ACROSS SPEAKERS

To analyze the data across speakers, a dataset was analyzed that included 58 glottal pulses from each of 9 speakers. Half of each speaker's 58 pulses was spoken under stress, while the other half was spoken under normal conditions. Eight measures were made on the glottal pulses; namely, amplitude (amp); opening slope (os); closing slope (cs); ratio of opening to closing slope (ratio); pitch; and three beta-function parameters, A (AA), B (BB), and C (CC).

The three independent variables were speaker (b1, b2, b3, g1, g2, g3, n1, n2, or n3), condition (cond70 or normal), and the interaction of speaker with condition (b1-cond70, b1-normal, b2-cond70, b2-normal, etc.). Each measure was a dependent variable. With this assignment to the variables, we performed an initial MANOVA to determine whether any significant difference existed in the means of any of the dependent variables across the independent variables.

With the MANOVA as confirmation that some significance existed, separate ANOVAs were performed, one for each dependent variable. Each ANOVA indicated whether any significant difference existed in the means across speakers. The following is a sample ANOVA:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	3.16E10	3.95E9	17.514	0.0001
cond	1	1.31E9	1.31E9	5.807	0.0163
cond * spkr	8	7.723E9	9.654E8	4.279	0.0001
Residual	504	1.137E11	2.255E8		

Dependent: cs

It shows that the dependent variable is cs (closing slope), and the most important information for our purpose is the "P-Value" on the far right of the table. If the "P-Value" for a given independent variable is less than "0.05," then that indicates significance at the 95-percent confidence level. In this table, significance is indicated for all three independent variables, spkr, cond, and the interaction of spkr and cond. Significance can be misleading, however. For example, a significant indicator for spkr in this table means simply that at least one speaker's mean closing slope varied significantly from the other speakers' mean closing slopes. It does not indicate they all had significantly different means. Similarly, the interaction significance indicates simply that the mean closing slope of at least one speaker-condition pair (e.g., b3-normal) varied significantly from the other speaker-condition pairs. To determine where the significance lies requires each instance of the independent variable to be analyzed with each other instance, e.g., a means comparison of one speaker with every other speaker. Since condition was the main focus of this analysis, the speaker and interaction analysis was not pursued extensively.

A significant indicator for condition is relevant. Since only two conditions (cond70 and normal) are considered, significance indicates the dependent-variable means of stressed (cond70) and normal speech varied significantly. This is what we have been looking for; that across speakers, the value of the dependent variable appears to be significantly different between stressed and normal. Unfortunately, a significant difference in the "mean" value across speakers does not necessarily mean that the dependent variable was significantly different for stressed and normal of each and every speaker. Also, a dependent variable that is significant for two different speakers might be significant in different ways. For example, closing slope might rise significantly for one speaker under stress, but lower significantly for another speaker under stress. Hence, each indicator of significance is useful, but not conclusive. The following will demonstrate that each lead must be tracked down to determine the true significance of each variable.

ANALYSIS PART II: WITHIN SPEAKERS

The "within" speaker analysis allowed use of a slightly broader data set. Since each speaker was analyzed individually, the number of samples for each speaker could be expanded, if such data were available. Whereas in the "across" speaker analysis, we used 58 samples from each speaker; and in the "within" speaker analysis, we used at least 58 samples for each speaker—and more, if the data were available. The only requirement was that the same number of stressed and normal samples be used for any one speaker.

There was only one independent variable, cond. The goal was to compare the mean values of the eight dependent variables across this one independent variable to determine which dependent variables, if any, were significant for each speaker. The dependent variables were the standard eight measures of the glottal pulse.

The following sample ANOVA is for speaker b3:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	13.158	13.158	62.069	0.0001
Residual	60	12.719	0.212		

Dependent: CC

It shows that CC, one of the beta-function parameters, was the dependent variable; and that it varied significantly when the independent variable switched from normal to cond70. The significance is indicated by the small "P-Value," below 0.05.

Before analyzing the dependent-variable significance for each speaker, a MANOVA was performed for each speaker. The MANOVA indicated whether any of the dependent variables varied significantly across condition. With a significant indicator here, we are justified to proceed with individual ANOVAs for each dependent variable.

Now that the basic analysis has been described, the results (Chapter 8) can be reviewed. This review analyzes the results in some detail; however, later paragraphs in Chapter 8 summarize the results.

CHAPTER 8
RESULTS

ACROSS SPEAKERS

The description of results will first cover MANOVAs, with a significant MANOVA justifying further analysis by ANOVAs on the dependent measures. These ANOVAs will be evaluated to determine what measures appear significant. Finally, the speaker-condition-value graphs will be reviewed for each dependent variable, providing a direct look at the means. These graphs indicate to what degree and in what manner the measures vary.

Before proceeding (for those who may be browsing through this report), the conclusions reached from the results will be briefly previewed here. In essence, the findings are as follows:

1. The MANOVAs indicate a significant difference in the means across speakers, justifying further analysis by separate ANOVAs on the individual measures.

2. The ANOVAs confirm that mean closing slope, pitch, AA, BB, and CC vary significantly between stressed and normal speech for these nine speakers. The mean values suggest that the closing slope falls under stress, the pitch rises approximately 4 Hz, and the glottal pulse narrows (as AA, BB, and CC rise).

3. The individual graphs of speaker-condition-value for each measure suggest that, although the measures of closing slope, pitch, AA, BB, and CC may be significantly different across speakers for condition, they are in many ways still speaker dependent. For example, pitch may rise for three speakers, lower for two, and show no significant difference between condition for three others.

4. These results are important because they suggest that some excitation signal measures are significant for determining stress across speakers. Yet, they are also important for indicating the remaining speaker-dependencies and the likely need to look at performing speaker-dependent analysis for effective stress detection.

First, the across-speaker MANOVAs are shown:

Type III MANOVA Table
Effect: spkr
S 8
M –0.500
N 247.500

	Value	F-Value	Num DF	Den DF	P-Value
Wilk's Lambda	0.040	33.536	64.000	2873.121	0.0001
Roy's Greatest Root	4.771				
Hotelling-Lawley Trace	6.729	52.069	64.000	3962.000	0.0001
Pillai Trace	1.977	20.673	64.000	4032.000	0.0001

Type III MANOVA Table
Effect: cond
S 1
M 3.000
N 247.500

	Value	F-Value	Num DF	Den DF	P-Value
Wilk's Lambda	0.848	11.154	8.000	497.000	0.0001
Roy's Greatest Root	0.180	11.154	8.000	497.000	0.0001
Hotelling-Lawley Trace	0.180	11.154	8.000	497.000	0.0001
Pillai Trace	0.152	11.154	8.000	497.000	0.0001

Type III MANOVA Table
Effect: cond * spkr
S 8
M –0.500
N 247.500

	Value	F-Value	Num DF	Den DF	P-Value
Wilk's Lambda	0.529	5.238	64.000	2873.121	0.0001
Roy's Greatest Root	0.462				
Hotelling-Lawley Trace	0.731	5.654	64.000	3963.000	0.0001
Pillai Trace	0.562	4.756	64.000	4032.000	0.0001

Each table analyzes whether any dependent variable significantly varied for one of the three independent variables: (1) spkr, (2) cond, and (3) the interaction of spkr and cond (cond*spkr). Up to four separate methods for computing the MANOVA were performed, and the results were included in each table. All of the tests showed significance, that is, "P-Value" below 0.05, which simply justifies our looking forward to performing separate analyses. The MANOVA looks for any significance across all the data, whereas each ANOVA looks for significance only on the one dependent variable. Therefore, the MANOVA is an important first step, and without it, one might be misled when performing the more specific ANOVAs.

Now that the MANOVAs have confirmed that some dependent variable means are significantly different for condition, ANOVAs must be performed for each dependent variable to determine which hold the significance. Each individual ANOVA will indicate whether a given measure of the glottal pulse has any significance. The following evaluations cover the eight respective ANOVAs, namely:

1. Amplitude—Across Speakers

Type III Sums of Squares

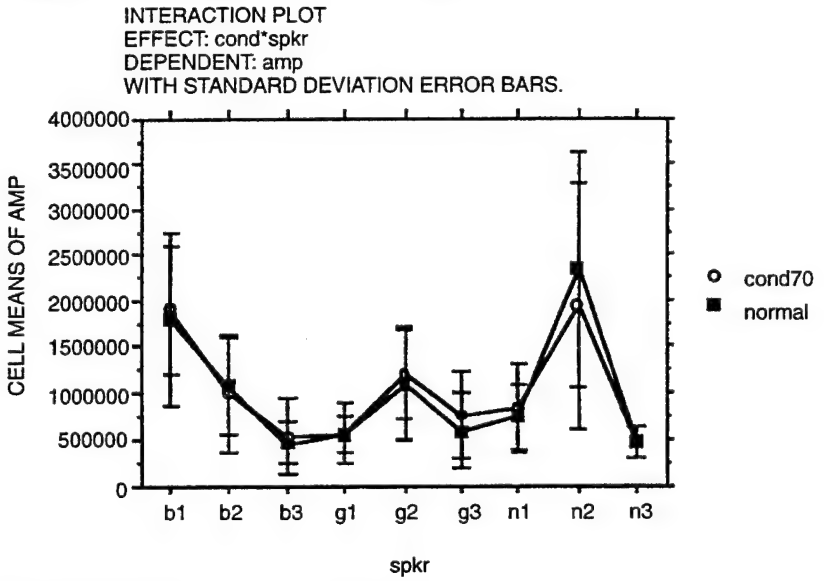
Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	1.688E14	2.11E13	52.256	0.0001
cond	1	1.181E10	1.181E10	0.029	0.8643
cond * spkr	8	3.217E12	4.021E11	0.996	0.4383
Residual	504	2.036E14	4.039E11		

Dependent: amp

This ANOVA suggests that (a) at least one speaker had a significantly different mean amplitude from the other speakers' mean amplitudes (P-Value = 0.0001, less than 0.05); (b) amplitude is not significantly different between stressed and normal speech across speakers (P-Value = 0.8643, not less than 0.05); and (c) no mean amplitude is present for any speaker-condition pair that is significantly different from the mean amplitudes of the other speaker-condition pairs (P-Value = 0.4383, not less than 0.05). In short, the only significance of mean amplitude is to distinguish speakers (or at least one speaker), but not to distinguish condition. From this analysis, one might want to explore how amplitude varies by speaker, if interested in using this measure for some other purpose than stress detection, perhaps for speaker identification. For stress detection, this study suggests amplitude is irrelevant.

Our intuition might suggest that when someone is under stress, their speech might get louder or perhaps they might speak more softly. If this were the case for most speakers, then one would expect to find a significant difference in amplitude between stressed and normal speech across speakers. Our results suggest this is not the case. The subjects under our analysis appear not to have raised or lowered their voices under stress. No justification exists for rejecting the null hypothesis: no significant difference occurs in amplitude, despite our possible intuition.

The following graph displays the amplitude values for stressed and normal speech for each speaker, thus giving us a visual picture of how amplitude varied in our study:



This graph is fairly easy to understand. The normal and stressed amplitude means are almost exactly the same for every speaker, especially when one allows for possible variance indicated by the standard deviation brackets. Clearly, amplitude will not be a likely place to find a good stress indicator in our data.

Although the ANOVA indicated significance for spkr, further analysis is required to confirm across how many speakers this significance lies. The ANOVA significance simply represents that at least one speaker has a significantly different amplitude mean. The graph serves nicely as a visual aid for intuitive confirmation, but can be misleading for reliable statistical analysis. A Tukey-Kramer analysis is a standard method for comparing the significance of means; in this case, each speaker's mean with each other speaker's mean and determining any significant difference. In the following table, a significant difference at the 5-percent level is indicated by an "s" to the right of the spkr versus spkr row. The more "s" marks down the right side of the Tukey-Kramer chart, the more widespread is the significance of the condition across speakers.

Tukey-Kramer
Effect: spkr
Dependent: amp
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
n3	b3	29526.9283.663E5		
	g1	86251.9333.663E5		
	g3	2.134E5	3.663E5	
	n1	3.271E5	3.663E5	
	b2	5.733E5	3.663E5	S
	g2	6.843E5	3.663E5	S
	b1	1.383E6	3.663E5	S
	n2	1.667E6	3.663E5	S
b3	g1	56725.0053.663E5		
	g3	1.838E5	3.663E5	
	n1	2.976E5	3.663E5	
	b2	5.438E5	3.663E5	S
	g2	6.548E5	3.663E5	S
	b1	1.354E6	3.663E5	S
	n2	1.637E6	3.663E5	S
g1	g3	127106.993.663E5		
	n1	2.409E5	3.663E5	
	b2	4.871E5	3.663E5	S
	g2	5.98E5	3.663E5	S
	b1	1.297E6	3.663E5	S
	n2	1.58E6	3.663E5	S
g3	n1	1.138E5	3.663E5	
	b2	359960.113.663E5		
	g2	4.709E5	3.663E5	S
	b1	1.17E6	3.663E5	S
	n2	1.453E6	3.663E5	S
n1	b2	2.462E5	3.663E5	
	g2	3.572E5	3.663E5	
	b1	1.056E6	3.663E5	S
	n2	1.34E6	3.663E5	S

b2	g2	1.11E5	3.663E5	
	b1	8.097E5	3.663E5	S
	n2	1.093E6	3.663E5	S
g2	b1	6.987E5	3.663E5	S
	n2	9.824E5	3.663E5	S
b1	n2	2.837E5	3.663E5	
S = Significantly different at this level.				

The Tukey-Kramer analysis suggests the significance of mean amplitude among the speakers extends beyond just one or two of the speakers. Many of the speakers, although not all, differ significantly in mean amplitude. This could be a useful result when considering speaker-identification applications. And, yet, one must note that changes in unmeasured factors, such as microphone distance from the lips, can greatly affect amplitude measurement. With this caveat, amplitude does appear significant for distinguishing speakers, although not for distinguishing stressed from normal speech.

2. Opening Slope

The following discussion gives the findings on whether mean opening slope varied among the independent variables, spkr, cond, and cond*spkr. The following ANOVA table presents data for opening slope across speakers:

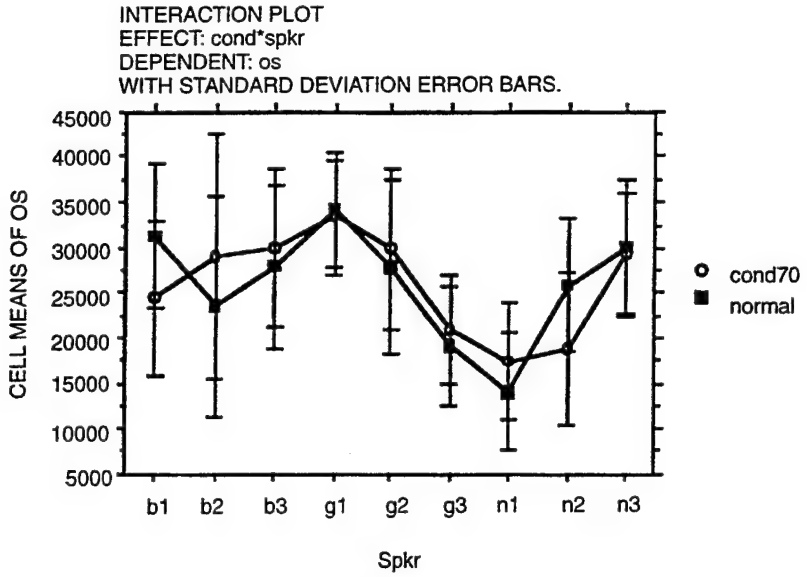
Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	1.441E10	1.801E9	25.526	0.0001
cond	1	1886644.118	1886644.118	0.027	0.8702
cond * spkr	8	2.212E9	2.765E8	3.918	0.0002
Residual	504	3.556E10	7.056E7		

Dependent: os

The results of this ANOVA suggest (1) at least one speaker's mean opening slope varies significantly from the other speakers' mean opening slopes (P-Value = 0.0001, less than 0.05); (2) the mean opening slope of stressed speech glottal pulses is not significantly different from the mean opening slope of normal glottal pulses (P-Value = 0.8702, not less than 0.05); and (3) the mean opening slope of at least one speaker-condition pair is significantly different from the mean opening slopes of other speaker-condition pairs (P-Value = 0.0002, less than 0.05).

The first two suggested conclusions are similar to our earlier amplitude results. As just discussed, we can note that opening slope is not a likely indicator of stressed speech, yet it may be useful for speaker identification. To see how widespread the significance of opening slope is across the speakers, the Tukey-Kramer analysis can be performed as we did for Amplitude. Before doing the Tukey-Kramer, to get an intuitive feel for why the ANOVA came out as it did, note the following:



This graph shows that the means for cond70 and normal for most of the speakers are close together, but not exactly the same. The graph suggests that perhaps, for some speakers, opening slope might be significant. The answer to this is given later on in this chapter, where speaker analysis is performed. Yet, apparently, the significance was not sufficiently widespread to show up in our ANOVA as generally significant across all speakers.

The graph also suggests a variability by speaker, since the means rise and fall across the x-axis. Note also that the condition means track the rise and fall of speaker variability quite well. This helps to explain why the ANOVA showed a significant interaction between condition and speaker. Since the ANOVA found significant speaker variability, a significant interaction suggests the condition followed the speaker variability.

The Tukey-Kramer analysis of the significance of mean opening slope for distinguishing speakers can now be shown:

Tukey-Kramer
Effect: spkr
Dependent: os
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
n1	g3	4241.176	4842.013	
	n2	6525.762	4842.013	S
	b2	10458.830	4842.013	S
	b1	12126.719	4842.013	S
	g2	13036.196	4842.013	S
	b3	13208.435	4842.013	S
	n3	13923.389	4842.013	S
	g1	17987.577	4842.013	S
g3	n2	2284.586	4842.013	
	b2	6217.654	4842.013	S

	b1	7885.543	4842.013	S
	g2	8795.020	4842.013	S
	b3	8967.259	4842.013	S
	n3	9682.213	4842.013	S
	g1	13746.401	4842.013	S
n2	b2	3933.067	4842.013	
	b1	5600.956	4842.013	S
	g2	6510.434	4842.013	S
	b3	6682.673	4842.013	S
	n3	7397.627	4842.013	S
	g1	11461.815	4842.013	S
b2	b1	1667.889	4842.013	
	g2	2577.367	4842.013	
	b3	2749.605	4842.013	
	n3	3464.559	4842.013	
	g1	7528.748	4842.013	S
b1	g2	909.478	4842.013	
	b3	1081.716	4842.013	
	n3	1796.670	4842.013	
	g1	5860.859	4842.013	S
g2	b3	172.239	4842.013	
	n3	887.192	4842.013	
	g1	4951.381	4842.013	S
b3	n3	714.954	4842.013	
	g1	4779.142	4842.013	
n3	g1	4064.189	4842.013	

S = Significantly different at this level.

This analysis suggests the significance may not be as widespread as was amplitude, but rather clustered. Four speakers (n1, g3, n2, and g1) are significantly different from the others. As with amplitude, opening slope appears to be a useful factor in distinguishing speakers, but not for distinguishing stressed vs. normal speech.

3. Closing Slope

Closing Slope is the first measure which we will find has a mean that varies significantly between stressed and normal speech across speakers. The ANOVA results are as follows:

Type III Sums of Squares

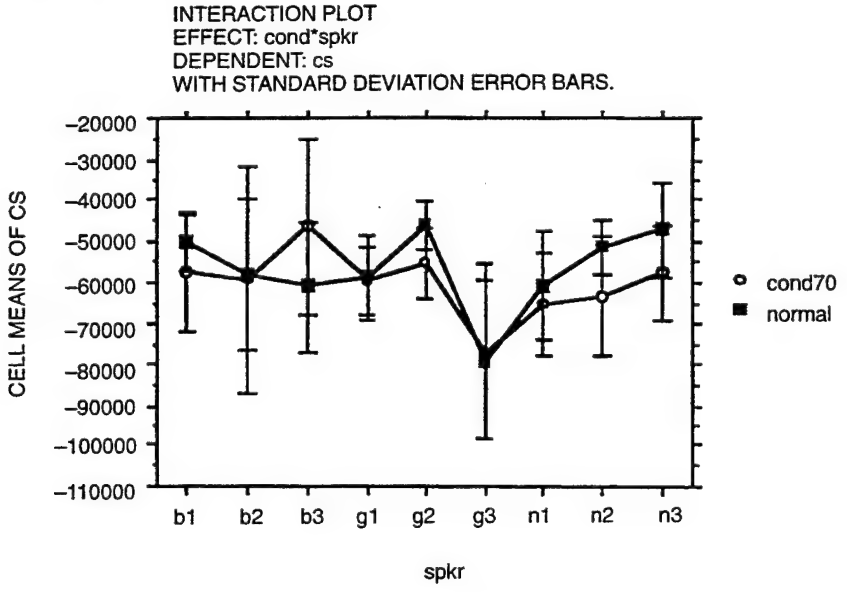
Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	3.16E10	3.95E9	17.514	0.0001
cond	1	1.31E9	1.31E9	5.807	0.0163
cond * spkr	8	7.723E9	9.654E8	4.280	0.0001
Residual	504	1.137E11	2.255E8		

Dependent: cs

The ANOVA table shows the closing slope to vary significantly for all independent measures. Specifically, the table suggests (1) at least one speaker's mean closing slope varies significantly from at least one other speaker's mean closing slope (P-Value = 0.0001, less than 0.05); (2) the mean closing slope of the stressed pulses varies significantly from the mean closing slope of the normal pulses (P-Value = 0.0163, less than 0.05); and (3) the mean closing slope of at least one speaker-condition pair varies significantly from the mean closing slope of some other speaker-condition pair (P-Value = 0.0001, less than 0.05).

Note that significance for an independent variable, such as speaker, signifies only that at least two values of that independent variable, i.e., two speakers, have means for this dependent variable, closing slope, that vary significantly. Further analysis must be performed to determine exactly how the significance is spread across the various values of the independent variable. Since condition has only two values (stressed or normal), significance for this independent variable is actually meaningful without additional analysis.

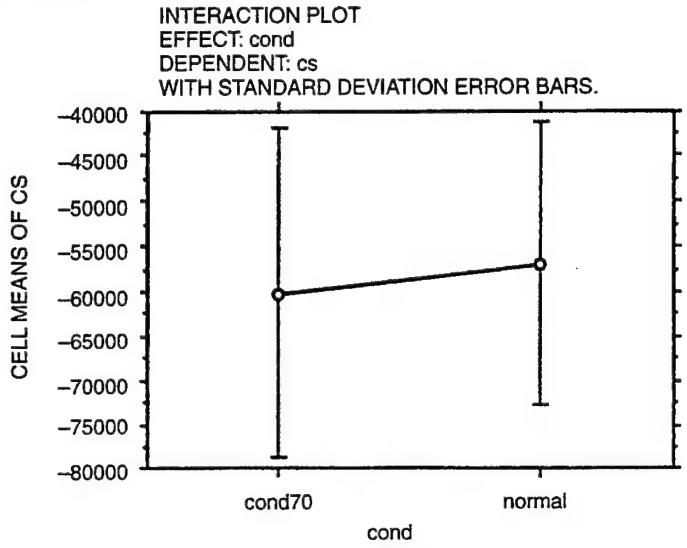
The following graph shows the closing slope mean values, providing an intuitive understanding of these ANOVA results:



Note from the graph that the mean closing slope value for stressed pulses (cond70) appears to be lower for the majority of speakers (b1, g2, n1, n2, and n3). Since the closing slope is negative, a lower closing slope means a steeper slope downward. This would correspond with a faster closing of glottis than with normal speech. The next section of this chapter will indicate if these

slope changes are significant within each speaker. Note also that for three of the speakers (b2, g1, and g3), mean closing slope appears the same, despite condition. Finally, for one speaker (b3), the closing slope appears to rise under stressed speech. Clearly, the effect on closing slope is different for different speakers; but the ANOVA table results suggest that for a majority of speakers, the effect is common enough to suggest this feature is significant across speakers.

The mean closing slope by condition can be plotted across speakers to see the difference the ANOVA found significant:



From this graph, the mean closing slope appears lower for stressed speech (cond70) across speakers, but not by much. The ANOVA found this significant, but not as significant as the variation of closing slope mean by speaker and speaker*cond70 interaction.

The variation of closing slope mean by speaker needs further analysis to determine how many speakers vary significantly from each other. Here is the Tukey-Kramer analysis:

Tukey-Kramer
Effect: spkr
Dependent: cs
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
g3	n1	15088.835	8657.031	S
	g1	18710.901	8657.031	S
	b2	19217.062	8657.031	S
	n2	20720.749	8657.031	S
	b1	24066.745	8657.031	S
	b3	24256.587	8657.031	S
	n3	25758.719	8657.031	S
	g2	27267.889	8657.031	S
n1	g1	3622.066	8657.031	
	b2	4128.227	8657.031	

	n2	5631.914	8657.031	
	b1	8977.910	8657.031	S
	b3	9167.752	8657.031	S
	n3	10669.884	8657.031	S
	g2	12179.055	8657.031	S
g1	b2	506.161	8657.031	
	n2	2009.848	8657.031	
	b1	5355.844	8657.031	
	b3	5545.686	8657.031	
	n3	7047.818	8657.031	
	g2	8556.988	8657.031	
b2	n2	1503.687	8657.031	
	b1	4849.683	8657.031	
	b3	5039.525	8657.031	
	n3	6541.657	8657.031	
	g2	8050.827	8657.031	
n2	b1	3345.996	8657.031	
	b3	3535.837	8657.031	
	n3	5037.970	8657.031	
	g2	6547.140	8657.031	
b1	b3	189.841	8657.031	
	n3	1691.974	8657.031	
	g2	3201.144	8657.031	
b3	n3	1502.133	8657.031	
	g2	3011.303	8657.031	
n3	g2	1509.170	8657.031	

S = Significantly different at this level.

The Tukey-Kramer analysis suggests the significance is not widespread. Two speakers appear significantly different from most of the others, namely, g3 and n1. The other speakers do not have means that vary significantly from each other. Thus, closing slope is probably not a reliable measure for distinguishing speakers.

4. Ratio of Opening Slope to Closing Slope

The ratio of the slopes was analyzed as a separate measure, but proved insignificant for distinguishing stressed speech. Here are the ANOVA results:

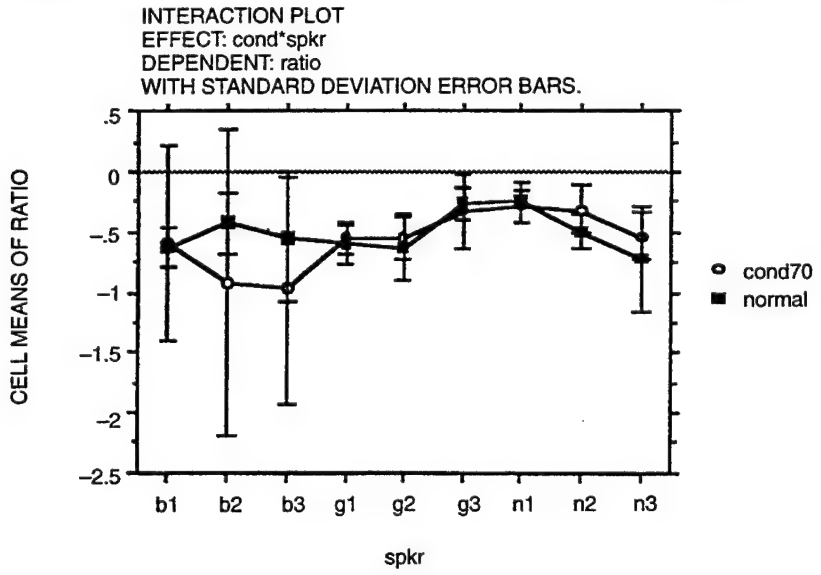
Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	13.560	1.695	7.305	0.0001
cond	1	0.357	0.357	1.539	0.2153
cond * spkr	8	6.746	0.843	3.634	0.0004
Residual	504	116.939	0.232		

Dependent ratio

The ANOVA table indicates (1) at least one speaker's mean ratio varied significantly from at least one other speaker's mean ratio (P-Value = 0.0001, less than 0.05); (2) the mean ratio for stressed speech did not vary significantly from the mean ratio for normal speech (P-Value = 0.2153, not less than 0.05); and (3) the mean for at least one speaker-condition pair varied significantly from at least one other speaker-condition pair (P-Value = 0.0004, less than 0.05).

Unfortunately, mean ratio does not vary significantly between stressed and normal speech. We can verify this by referring to the graph of mean ratio values by speaker-condition:



The graph shows that for most speakers (b1, g1, g2, g3, and n1), the stressed and normal means for ratio are almost the same. This helps to explain why the ANOVA did not find ratio different for condition. At the same time, the mean ratios are different among a few speakers, which explains why the ANOVA found that the means varied significantly by speaker.

The following Tukey-Kramer analysis shows the mean ratio by speaker:

Tukey-Kramer
Effect: spkr
Dependent: ratio
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
b3	b2	0.082	0.278	
	n3	0.128	0.278	
	b1	0.151	0.278	
	g2	0.167	0.278	
	g1	0.174	0.278	
	n2	0.347	0.278	S
	g3	0.464	0.278	S
	n1	0.489	0.278	S
b2	n3	0.046	0.278	
	b1	0.069	0.278	
	g2	0.085	0.278	
	g1	0.092	0.278	
	n2	0.265	0.278	
	g3	0.382	0.278	S
	n1	0.407	0.278	S
n3	b1	0.023	0.278	
	g2	0.039	0.278	
	g1	0.046	0.278	
	n2	0.219	0.278	
	g3	0.336	0.278	S
	n1	0.361	0.278	S
b1	g2	0.017	0.278	
	g1	0.023	0.278	
	n2	0.196	0.278	
	g3	0.313	0.278	S
	n1	0.338	0.278	S
g2	g1	0.006	0.278	
	n2	0.180	0.278	
	g3	0.297	0.278	S
	n1	0.322	0.278	S
g1	n2	0.173	0.278	
	g3	0.291	0.278	S

	n1	0.316	0.278	S
n2	g3	0.117	0.278	
	n1	0.142	0.278	
g3	n1	0.025	0.278	

S = Significantly different at this level.

Although several "s" symbols are immediately adjacent to the right of the table, they are associated with only a few speakers. In essence, two speakers (g3 and n1) have means that vary significantly from the others. So ratio may be significant for distinguishing some speakers from other speakers, but this difference is not widely significant, that is, many speakers cannot be distinguished by using this measure.

5. Pitch

Pitch is a measure that one might intuitively expect to vary significantly across speakers and across conditions. In fact, the following ANOVA table suggests our intuition is correct:

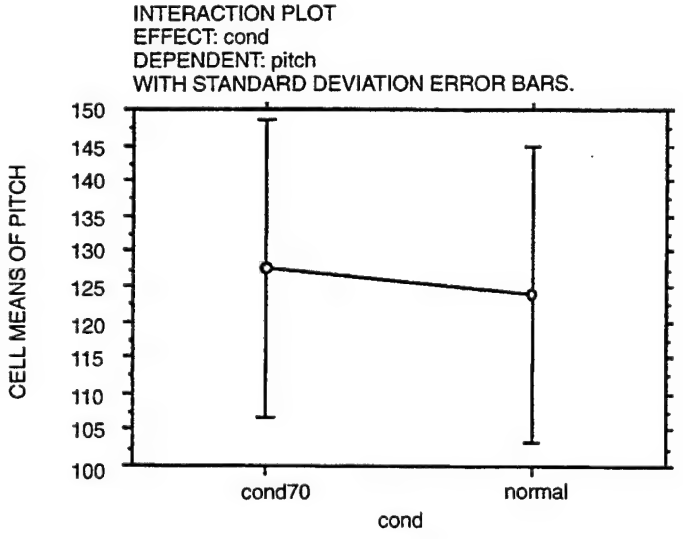
Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	169209.648	21151.206	264.467	0.0001
cond	1	1646.224	1646.224	20.584	0.0001
cond * spkr	8	14544.552	1818.069	22.732	0.0001
Residual	504	40308.345	79.977		

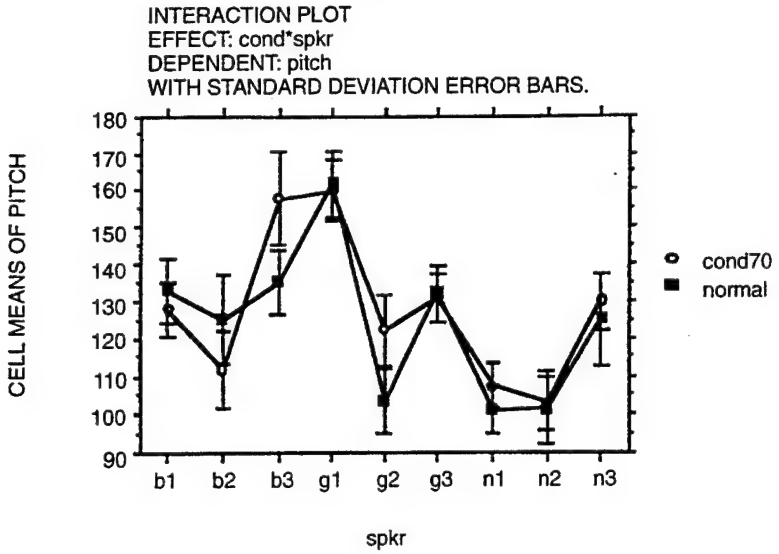
Dependent pitch

The table suggests (1) at least one speaker's mean pitch varies significantly from at least one other speaker's mean pitch (P-Value = 0.0001, less than 0.05); (2) mean pitch of stressed speech varies significantly from mean pitch of normal speech (P-Value = 0.0001, less than 0.05); and (3) at least one speaker-condition pair mean pitch varies significantly from at least one other speaker-condition pair mean pitch.

First, the following plot of mean pitch by condition shows how the mean pitch varies by condition:



The graph suggests mean pitch rises perhaps 3 Hz from normal speech to stressed speech (cond70). Three Hertz is not much, so mean pitch variance should be explored further by looking at a graph of the mean pitch values by speaker and condition, as follows:



This graph shows that pitch variances are speaker-dependent. For example, pitch lowered for stressed speech of two speakers (b1 and b2), rose for four speakers (b3, g2, n1, and n3), and stayed the same for two speakers (g1 and g3). The following within-speaker analysis will reveal more about whether these variances in means are significant for each speaker separately. For now, pitch appears to be significant, but it does not vary in exactly the same way for all speakers.

The wandering lines on the graph provide an intuitive confirmation of the other ANOVA table findings, namely, that pitch means vary by speaker, and that the variances in mean by condition appear correlated with speaker. The following Tukey-Kramer analysis tells more about how widespread the speaker significance is:

Tukey-Kramer
Effect: spkr
Dependent: pitch
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
n2	n1	2.069	5.155	
	g2	10.603	5.155	S
	b2	16.017	5.155	S
	n3	25.017	5.155	S
	b1	28.190	5.155	S
	g3	29.155	5.155	S
	b3	44.069	5.155	S
	g1	58.207	5.155	S
n1	g2	8.534	5.155	S
	b2	13.948	5.155	S
	n3	22.948	5.155	S

	b1	26.121	5.155	S
	g3	27.086	5.155	S
	b3	42.000	5.155	S
	g1	56.138	5.155	S
g2	b2	5.414	5.155	S
	n3	14.414	5.155	S
	b1	17.586	5.155	S
	g3	18.552	5.155	S
	b3	33.466	5.155	S
	g1	47.603	5.155	S
b2	n3	9.000	5.155	S
	b1	12.172	5.155	S
	g3	13.138	5.155	S
	b3	28.052	5.155	S
	g1	42.190	5.155	S
n3	b1	3.172	5.155	
	g3	4.138	5.155	
	b3	19.052	5.155	S
	g1	33.190	5.155	S
b1	g3	0.966	5.155	
	b3	15.879	5.155	S
	g1	30.017	5.155	S
g3	b3	14.914	5.155	S
	g1	29.052	5.155	S
b3	g1	14.138	5.155	S

S = Significantly different at this level.

In this case, the number of "s" markings on the right side of the Tukey-Kramer analysis confirm that the significance of pitch is widespread across speakers. That is, most speakers have a mean pitch that varies significantly from the mean pitch of most other speakers. One would expect from these results that pitch would be a useful measure of speaker identity.

6. Beta-Function Measure: AA

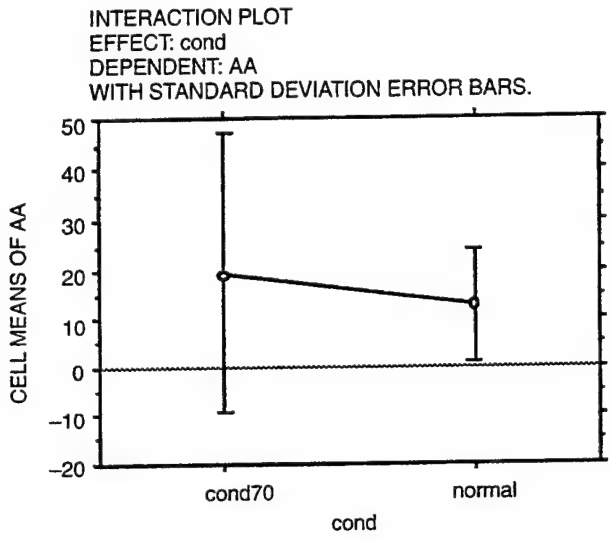
The beta-function measure, AA, proves to be a useful measure for distinguishing stress, as indicated by the following ANOVA table:

Type III Sums of Squares

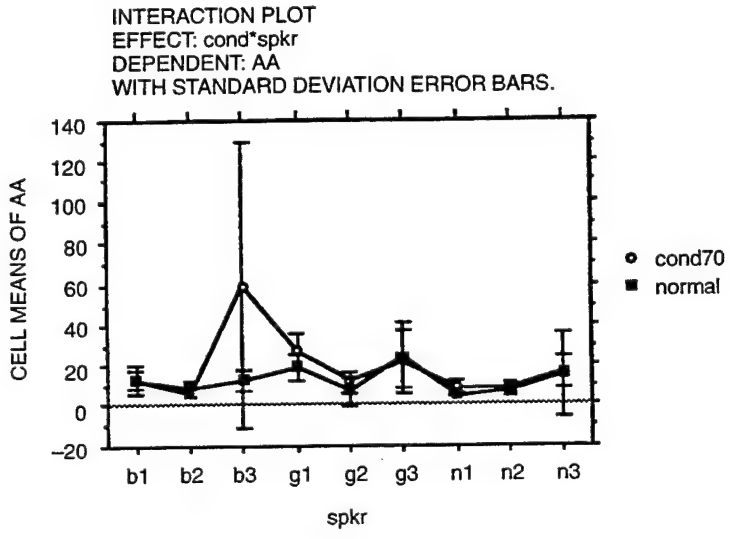
Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	45064.383	5633.048	16.198	.0001
cond	1	5838.504	5838.504	16.789	.0001
cond * spkr	8	26290.181	3286.273	9.450	0.0001
Residual	504	175269.751	347.757		

Dependent: AA

The ANOVA P-Values for all three independent variables suggest AA deserves further analysis. First, the following graph of mean AA values by condition shows the direction in which the mean AA varies significantly by condition:



The graph shows that AA rises across speakers under stress, from a value of 13 to around 18. The next graph shows how the AA values vary by speaker under stress:



However, this graph is difficult to interpret due to the large variance found in the cond70 values of b3. So, we will have to wait for the within speaker analysis to see for how many speakers the value AA is significant for distinguishing stressed speech.

The following Tukey-Kramer analysis aids in determining how AA varies by speaker:

Tukey-Kramer
Effect: spkr
Dependent: AA
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
n1	b2	0.685	10.750	
	n2	0.840	10.750	
	g2	2.614	10.750	
	b1	6.181	10.750	
	n3	8.786	10.750	
	g1	16.207	10.750	S
	g3	16.242	10.750	S
	b3	29.169	10.750	S
b2	n2	0.174	10.750	
	g2	1.949	10.750	
	b1	5.516	10.750	
	n3	8.120	10.750	
	g1	15.542	10.750	S
	g3	15.576	10.750	S
	b3	28.503	10.750	S
n2	g2	1.775	10.750	
	b1	5.342	10.750	
	n3	7.946	10.750	
	g1	15.367	10.750	S
	g3	15.402	10.750	S
	b3	28.329	10.750	S
g2	b1	3.567	10.750	
	n3	6.171	10.750	
	g1	13.593	10.750	S
	g3	13.627	10.750	S
	b3	26.554	10.750	S
b1	n3	2.604	10.750	
	g1	10.026	10.750	
	g3	10.060	10.750	

	b3	22.987	10.750	S
n3	g1	7.421	10.750	
	g3	7.456	10.750	
	b3	20.383	10.750	S
g1	g3	0.035	10.750	
	b3	12.962	10.750	S
g3	b3	12.927	10.750	S

S = Significantly different at this level.

This again shows a cluster of speakers without widespread variance across all speakers. Three speakers (g1, g3, and b3) apparently have distinctly different AA mean values from the other speakers. Yet, AA may not be a measure for easily distinguishing most speakers.

7. Beta-Function Measure: BB

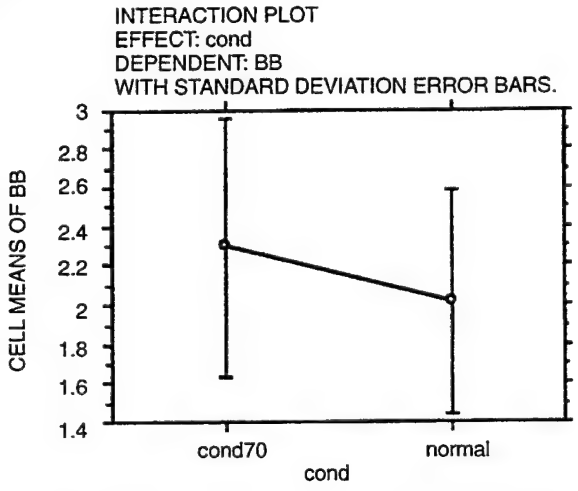
Here, mean BB values are a highly significant measure of condition. The following ANOVA table demonstrates this significance:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	101.863	12.733	78.727	.0001
cond	1	10.592	10.592	65.493	.0001
cond * spkr	8	15.104	1.888	11.673	.0001
Residual	504	81.514	0.162		

Dependent: BB

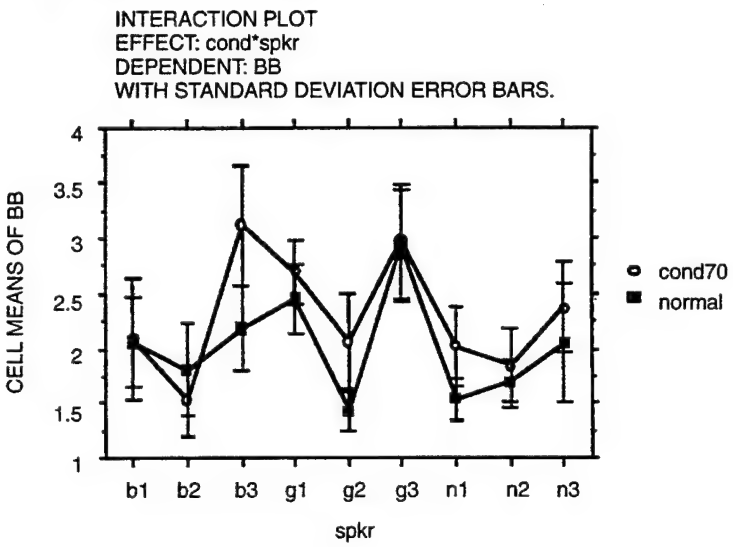
The BB P-Values confirm high significance for condition, as well as the other independent variables. The following graph shows how BB varies by condition:



The mean BB value is higher for stressed speech (cond70) across speakers. BB is a parameter in the beta function that controls the rising slope of the beta function. As BB rises, the initial beta function opening slope lowers. If the beta function pulses are normalized in amplitude, as

was done here, then a rising BB value for stress suggests the following. The pulse is either narrowed (if CC rises) or the pulse is shifted toward the closing slope (if CC lowers). As shown below, CC appears to rise with stressed speech, which suggests the pulses are narrowing under stress.

To get an intuitive feel for how widespread the BB rise is across speakers, see the following graph of BB values by speaker and condition:



The graph shows that BB appears to vary by speaker and that the mean values for stressed and normal appear correlated with the speakers. More importantly, for determining condition, we find that for six speakers, BB appears to rise under stress, whereas BB appears lower for one speaker(b2) under stress, and unchanged for two speakers (b1 and g3). Although BB was not a consistent factor for all speakers, it does appear to be more consistent than any other measure across the greatest number of speakers. If one were forced to rely on one measure to determine stressed speech across all speakers, BB appears to be the choice. Fortunately, such a choice may not be necessary, if we can develop systems that are speaker-specific. For such speaker-specific systems, we may have more reliability. Speaker-specific possibilities are analyzed in the following discussion of significant measures within speaker.

The Tukey-Kramer analysis shown here demonstrates the widespread significance of BB for speaker identification:

Tukey-Kramer
Effect: spkr
Dependent: BB
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
b2	g2	0.074	0.232	
	n2	0.098	0.232	
	n1	0.112	0.232	
	b1	0.407	0.232	S
	n3	0.544	0.232	S
	g1	0.903	0.232	S

	b3	0.978	0.232	S
	g3	1.283	0.232	S
g2	n2	0.024	0.232	
	n1	0.038	0.232	
	b1	0.332	0.232	S
	n3	0.469	0.232	S
	g1	0.829	0.232	S
	b3	0.904	0.232	S
	g3	1.209	0.232	S
n2	n1	0.014	0.232	
	b1	0.309	0.232	S
	n3	0.446	0.232	S
	g1	0.805	0.232	S
	b3	0.880	0.232	S
	g3	1.185	0.232	S
n1	b1	0.294	0.232	S
	n3	0.431	0.232	S
	g1	0.790	0.232	S
	b3	0.866	0.232	S
	g3	1.171	0.232	S
b1	n3	0.137	0.232	
	g1	0.496	0.232	S
	b3	0.571	0.232	S
	g3	0.876	0.232	S
n3	g1	0.359	0.232	S
	b3	0.434	0.232	S
	g3	0.739	0.232	S
g1	b3	0.075	0.232	
	g3	0.380	0.232	S
b3	g3	0.305	0.232	S

S = Significantly different at this level.

The large number of "s" marks to the right of the table show that mean BB value is significantly different for almost every speaker. BB would be a useful cue for a speaker-identification system.

8. Beta-Function Parameter: CC

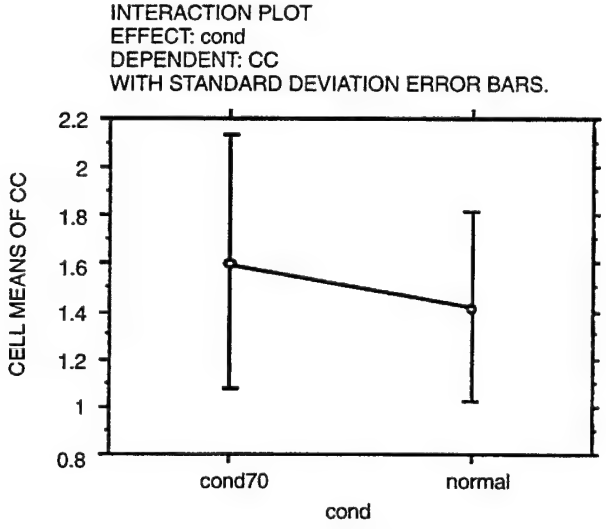
So far, beta-function parameters AA and BB have proved significant, which leaves CC for analysis. As with AA and BB, the ANOVA table for CC suggests that it, too, is significant for determining condition as well as speaker:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
spkr	8	50.077	6.260	62.386	0.0001
cond	1	4.450	4.450	44.350	0.0001
cond * spkr	8	12.033	1.504	14.990	0.0001
Residual	504	50.570	0.100		

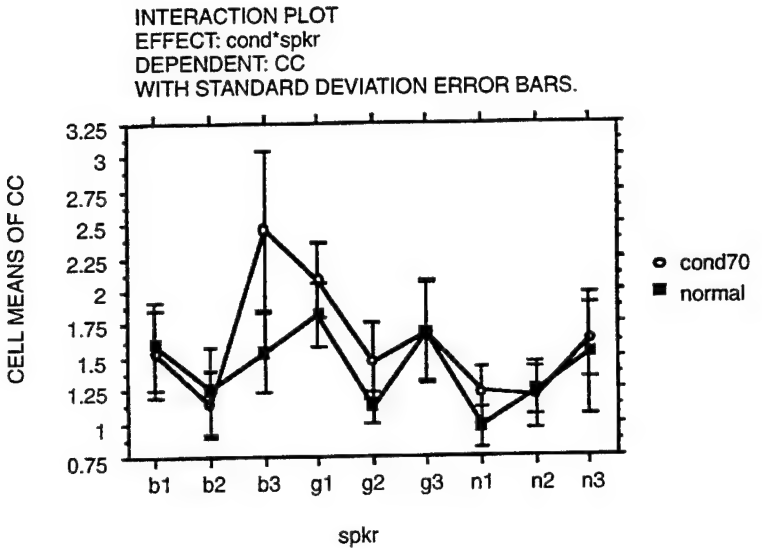
Dependent: CC

As the P-Values indicate, mean CC values vary significantly for all three independent variables. First, to see how it varies by condition, refer to the following plot of mean CC values by condition:



The graph suggests that CC rises under stress. CC controls the closing slope of the beta function, and a higher CC value indicates a lower slope at the end of the pulse. AA and BB appear to rise under stress; and the indication that CC also rises, confirms the earlier hypothesis that the pulses are narrowing under stress. The narrowing is not dramatic, but it does appear significant across the majority of speakers.

The following graph of CC values by speaker and condition indicates how many speakers appear to show a significant CC value change under stress:



The graph suggests that CC rises significantly for at least four speakers (b3, g1, g2, and n1) and appears unchanged for the other four. We will determine more specifically whether this measure is significant for each speaker when examining the measures within speaker. In addition, the graph suggests that BB varies significantly by speaker and that the mean condition values are correlated with speaker, as suggested by the ANOVA table.

Next, the Tukey-Kramer analysis shows for how many speakers the CC value is significantly different:

Tukey-Kramer
Effect: spkr
Dependent: CC
Significance level: 0.05

	Vs.	Diff.	Crit. diff.	
n1	b2	0.101	0.183	
	n2	0.125	0.183	
	g2	0.196	0.183	S
	b1	0.467	0.183	S
	n3	0.472	0.183	S
	g3	0.588	0.183	S
	g1	0.849	0.183	S
	b3	0.901	0.183	S
b2	n2	0.024	0.183	
	g2	0.095	0.183	
	b1	0.366	0.183	S
	n3	0.372	0.183	S
	g3	0.487	0.183	S
	g1	0.748	0.183	S
	b3	0.800	0.183	S

n2	g2	0.071	0.183	
	b1	0.342	0.183	S
	n3	0.348	0.183	S
	g3	0.463	0.183	S
	g1	0.724	0.183	S
	b3	0.776	0.183	S
g2	b1	0.271	0.183	S
	n3	0.276	0.183	S
	g3	0.392	0.183	S
	g1	0.653	0.183	S
	b3	0.705	0.183	S
b1	n3	0.005	0.183	
	g3	0.121	0.183	
	g1	0.382	0.183	S
	b3	0.434	0.183	S
n3	g3	0.116	0.183	
	g1	0.377	0.183	S
	b3	0.428	0.183	S
g3	g1	0.261	0.183	S
	b3	0.313	0.183	S
g1	b3	0.052	0.183	

S = Significantly different at this level.

As the "s" marks indicate, the significance of the CC value for distinguishing speakers is widespread. Most speakers analyzed had a CC value significantly different from the other speakers. Thus, CC appears to be a useful measure for speaker identification.

WITHIN SPEAKERS

The following discussion shifts our focus from across speakers to within speakers. This shift means that measures will no longer be sought that are effective for identifying stress across all speakers. Instead, a separate analysis will be done for each speaker; and each analysis will try to determine which measures are significant for identifying the stressful speech of this particular speaker.

Why the shift in focus to within speaker analysis? We would rather have found measures that were significant and reliable across all speakers for identifying stress. With such measures, a speaker-independent stress identification system might be developed. In fact, in our across speaker analysis, we found several parameters (closing slope, pitch, AA, BB, and CC) that are significant across a majority of speakers; however, the graphs confirmed that significant speaker variability remains.

We may be able to achieve more reliable and accurate stress identification, if different measures are used for different speakers. A within speaker analysis must be done to best determine

what measures work best for what speakers. We may find that for some speakers, none of the measures analyzed appears reliable. Yet the goal is to combine the best measures across speakers and then take advantage of speaker-specific measures, where possible, to best identify stress. For these reasons, within speaker analysis is pursued.

As a reminder at this point, our dataset is expanded. For across speaker analysis, 58 pulses were used for each speaker. The same number had to be used for each speaker, since statistics weighted to favor any particular speaker were not wanted. Now we are not so limited. Our statistical analysis evaluated pulses of each speaker individually; and so, the number of pulses for each speaker need not be the same. In fact, for within speaker analysis, the more pulses the better.

Due to artifacts in our semiautomatic data-collection process, different numbers of pulses were recovered for different speakers. The minimum number for any one speaker was 58. For the within speaker analysis, as many pulses were used for each speaker as our process allowed. The greatest number of pulses studied for any one speaker was 96.

Since only condition is being analyzed, the ANOVAs are much simpler and easier to understand than in the across speaker analysis. For this reason, we will only recount one speaker's analysis in detail; then we will summarize the results of all the other analyses in a chart. First, speaker b1 will be used as a sample case. This will demonstrate what the ANOVAs and related charts look like.

Sample Within Speaker Analysis: b1. Each speaker analysis begins with a MANOVA, telling us whether any significant variance exists among any of the mean values of the eight measures taken as a whole. For all speakers, except one, the MANOVA analyses showed significance. For example, here is the MANOVA for speaker b1:

Type III MANOVA Table
Effect: cond

S 1
M 3.000
N 42.500

	Value	F-Value	Num DF	Den DF	P-Value
Wilk's Lambda	.530	9.628	8.000	87.000	0.0001
Roy's Greatest Root	.885	9.628	8.000	87.000	0.0001
Hotelling-Lawley Trace	.885	9.628	8.000	87.000	0.0001
Pillai Trace	.470	9.628	8.000	87.000	0.0001

Here, four different types of MANOVA computations are shown, which will not be described in detail. The main point is that a P-Value of less than 0.05 indicates some significant variance exists. All four computations indicate significant variance. Now we are justified in looking at the individual ANOVAs for each measure.

Each ANOVA for a within speaker analysis has one independent variable, condition; and one dependent variable, that is, one of the eight measures. The ANOVA simply determines whether the mean value for the measure, for example, amplitude, varies significantly between stressed and normal speech for the given speaker. Here is the amplitude ANOVA for speaker b1:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	1.252E12	1.252E12	1.782	0.1851
Residual	94	6.602E13	7.024E11		

Dependent: amp

The amplitude ANOVA for speaker b1 shows a P-Value of 0.1851, which is not sufficient to reject the null hypothesis that the amplitude variance does not vary significantly between stressed and normal speech. In short, amplitude is not a significant measure for distinguishing b1's stressed speech, because the P-Value is not less than 0.05.

What about other measures? The following ANOVAs show b1's opening and closing slope measures:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	1.285E9	1.285E9	20.391	0.0001
Residual	94	5.924E9	6326093.39		

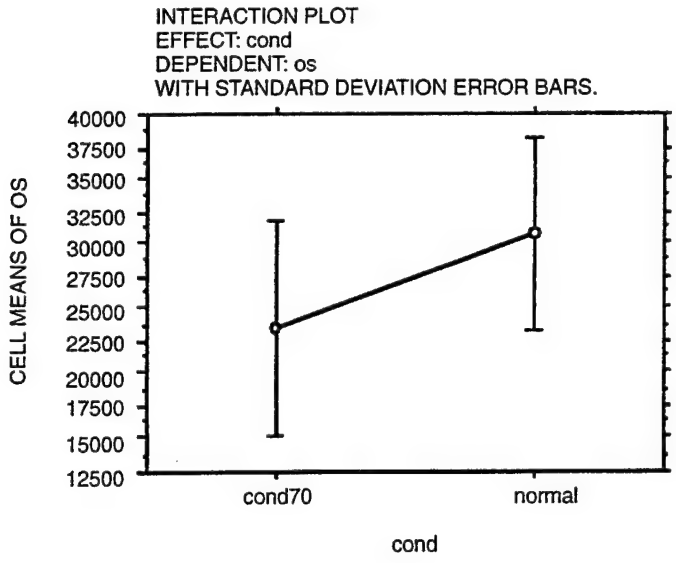
Dependent: os

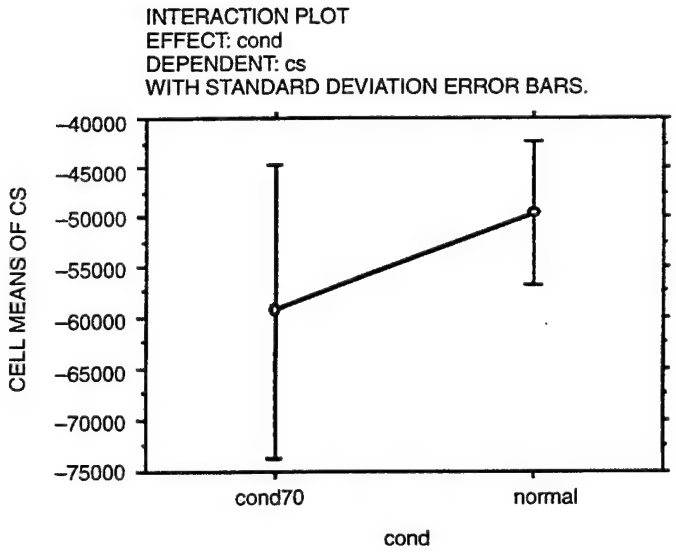
Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	2.214E9	2.214E9	16.935	0.0001
Residual	94	1.229E10	1.307E8		

Dependent: cs

The ANOVAs suggest that both opening and closing slopes are significant measures of stressed speech for speaker b1. To see how the slopes vary under stress, we can look at plots of the opening slope and closing slope means by condition for speaker b1:





The graphs show that both opening slope and closing slope lower for stressed speech. As discussed in Chapter 4, Quantitative Measures, the opening and closing slopes were measured at the midpoint of each half of the pulse, that is, halfway up the pulse and halfway down the pulse. Whereas AA, BB, and CC provide a global view of the pulse, our opening and closing slope measures provide pinpoint slope measurements at the rising and closing midpoints. So, the opening and closing slopes may simply reflect local changes in the shape of the pulses near the midpoints. The following discussion will show that global changes in pulse shape, measured by AA, BB, and CC, do not always correlate with local changes in the pulse shape caught by opening slope and closing slope.

At this point, we do not know if the opening and closing slope measurements are correlated with the global pulse measurements for speaker b1. However, this coverage will soon give an intuitive insight to it. For now, the ratio and pitch measures are presented for speaker b1:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	0.351	0.351	1.556	0.2154
Residual	94	21.234	0.226		

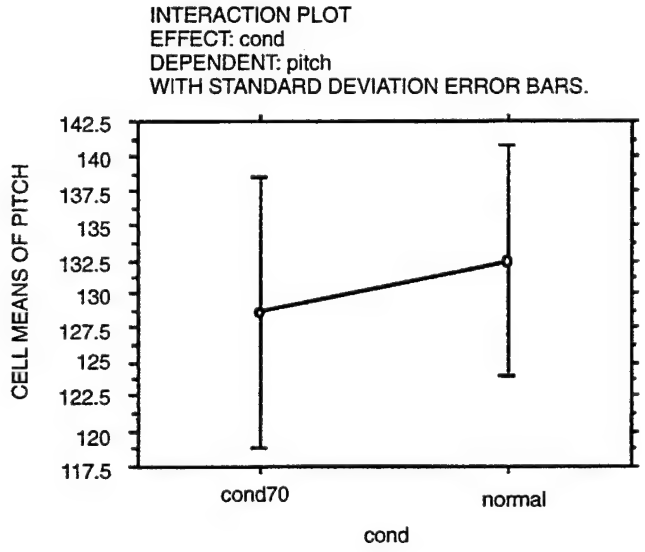
Dependent ratio

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	337.500	337.500	4.046	0.0471
Residual	94	7840.458	83.409		

Dependent pitch

These ANOVAs suggest ratio is not a significant measure with a P-Value of 0.2154, well above 0.05; but pitch is just barely significant at a P-Value of 0.0471. The following graph plots the pitch mean values for stressed and normal speech to show how pitch varies under stress for speaker b1:



The result is interesting; the pitch mean values suggest the pitch of speaker b1 actually lowers under stress by 3-4 Hz. And, although this is interesting, it may not be as reliable as other measures, since pitch was only marginally significant for this speaker.

Finally, the beta-function measures, AA, BB, and CC, are presented. For speaker b1, the ANOVAs determining the significance of these measures come out as follows:

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	67.712	67.712	1.320	0.2535
Residual	94	4820.882	51.286		

Dependent AA

Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	0.110	0.110	0.369	0.5451
Residual	94	28.109	0.299		

Dependent BB

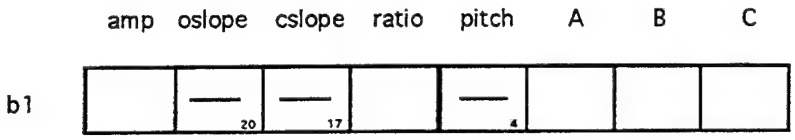
Type III Sums of Squares

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
cond	1	0.015	0.015	0.095	0.7591
Residual	94	14.636	0.156		

Dependent CC

The P-Values for these three parameters are all too high, 0.2535, 0.5451, and 0.7591, which are well above 0.05; so no reason exists to reject the null hypothesis. None of the mean values of AA, BB, or CC vary significantly between stressed and normal speech for this speaker, b1.

If the results of our ANOVA analysis for speaker b1 were to be summarized, the chart might look like this:



The chart shows one row assigned to speaker b1, where each column entry represents one of the eight measures. If a measure was found significant, a "+" or a "-" entry would appear in the corresponding column. For example, under opening slope (oslope), a "-" sign indicates that mean opening slope varied significantly between stressed and normal speech for this speaker, b1. A "-" means the mean varied significantly downward, going from normal to stressed speech. In other words, the mean value was lower for stressed speech than for the normal speech of this speaker. Similarly, a "+" sign would indicate the mean was higher for stressed speech. For example, the "-" sign under pitch indicates that the mean pitch value for speaker b1 was lower for stressed speech than for normal speech.

The chart contains one more piece of information, the relative degree of significance of the measure. This is reflected by the small number in the corner of each "+" or "-" entry. The number is the F-ratio number from the ANOVA for that measure and that speaker. The larger the F-Value, the more significant was the measure. For example, speaker b1 shows an F-Value of 20 for opening slope, 17 for closing slope, and 4 for pitch. This indicates that mean values for the

measures opening slope and closing slope varied by much greater amounts between stressed and normal speech than did the mean pitch. In fact, this piece of information reflects what we saw above when looking at the ANOVA's tables; that is, pitch was only marginally significant for speaker b1.

The summary chart provides the basic information from the ANOVA tables and mean plots. The advantage to the summary chart is that it provides the information in a much more compact, easy-to-read form. The chart contains only information for speaker b1; however, similar charts can be built for all speakers and the information can then be combined into one large matrix. This will allow review of the within-speaker results for each speaker, without having to tediously look at every ANOVA table and means plot.

Within Speaker Analysis: Summary Table

To review the within speaker analysis of the means variance of the eight measures across stressed versus normal speech, we developed a summary table. This matrix, described in the previous paragraph, identifies which measures were significant for each speaker. The significance is identified by a "+" or "-" marker, where "+" indicates the mean for that measure rose for stressed speech for that speaker; and a "-" indicates the mean fell for such stressed speech. Remember that the analyses were done "within" speaker, meaning that if a measure was found significant, it was found significant only for that speaker, not necessarily for all speakers. This is why a row is included for each speaker.

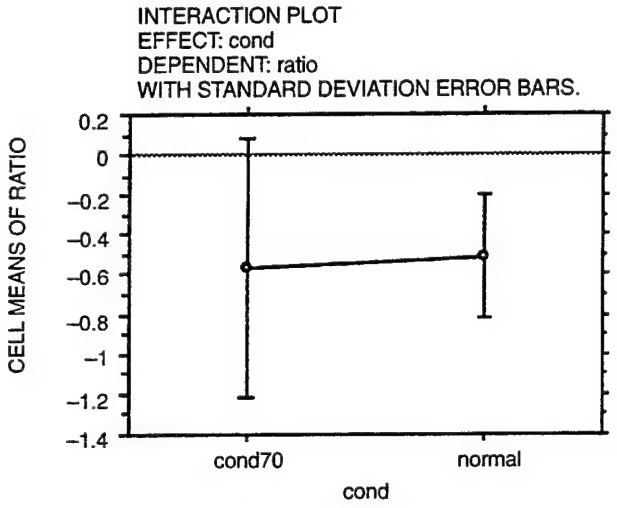
Here is the summary table for within speaker analysis:

	amp	oslope	cslope	ratio	pitch	A	B	C
b1		−20	−17		−4			
b2				−4	−20	−6	−8	
b3			+8	−4	+57	+14	+55	+62
g1				+5		+13	+9	+13
g2			−28		+84	+12	+70	+44
g3								
n1			−5		+12	+45	+57	+43
n2		−10	−15	+12			+4	
n3			−12	+4			+9	

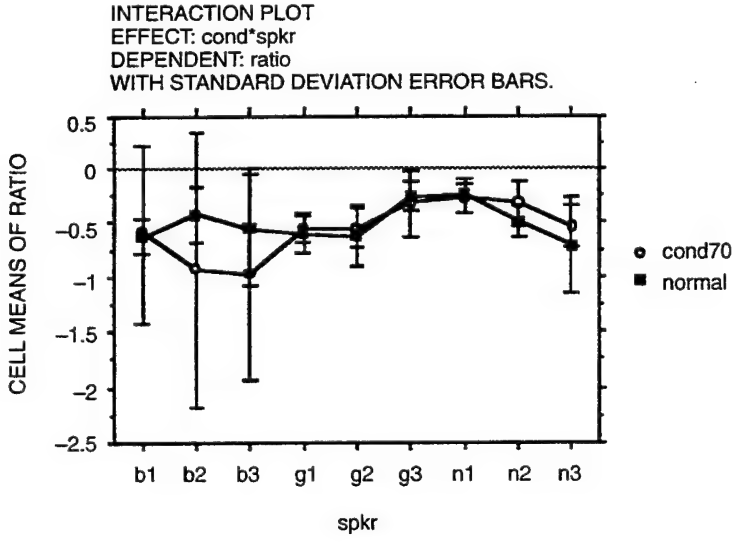
This table contains quite a bit of information that can be reviewed in different ways. For example, one can pick a speaker row and then look across the columns to see which measures were significant for that speaker. As another method for reviewing the data, one can pick a column and see for which speakers (and for how many speakers) this measure varied significantly. This discussion takes the latter approach, looking at each measure by column and reviewing its significance over the speakers. Before doing this, remember the difference between this type of analysis and the across speaker analysis performed earlier. In addition, a few general observations will be noted on this summary table before discussing the details.

So, if the summary table gives us the significance of a measure for each speaker, is that the same thing as our earlier analysis, which tried to determine if a measure was significant across all of the speakers? For example, if the summary table shows that a measure, such as ratio, was significant for a majority of speakers, does that mean it should have been significant in our earlier analysis; that is, when we tried to see if ratio was significant across all of the speakers? The answer to these questions is no. It is true that a measure that was found significant for identifying stress across speakers should show up as significant for many speakers in our within speaker analysis. The converse is not necessarily true. And, in fact, ratio, for example, appears significant in the summary table for five of the nine speakers. However, ratio was not found significant as a measure across speakers. Why is this?

This graph shows ratio means across all speakers:



The graph shows that the means do not differ by much, and easily seen is why the across speaker analysis did not find the mean ratio a significant measure. Yet, if we look at the plot of mean ratio by speaker and condition, the means may be significantly different for some speakers:



From looking at this graph, a different mean ratio (perhaps significant) appears to occur between stressed and normal speech for four of the speakers (b2, b3, n2, and n3). In fact, our summary table suggests ratio is also significant for g1, although the above plot does not easily give this visual impression. Why is this? From the graph, the means clearly vary by speaker and they are correlated with speaker. However, the means vary in different ways for different speakers (sometimes going up for stressed speech and sometimes going down). Consequently, the means come out almost the same when we look at the mean ratio for stressed and normal speech and ignore all of these speaker differences.

Another notable difference occurs between across speaker analysis versus within speaker analysis; that is, the latter contains only information on which measures are significant for distinguishing stressed versus normal speech. The across speaker analysis also contained information about which measures are useful for distinguishing speakers (for speaker identification applications) and for distinguishing speaker-condition pairs. Consequently, the within speaker analysis is much simpler and all of its information relates directly to our subject of interest, measuring stress in speech.

Before proceeding to detailed analysis of the summary table, we can make a few observations. First, if we look across the row for speaker g3, we see that no measure was found significant. In other words, none of our measures would be helpful for identifying whether g3 was under stress. This finding is in some ways unfortunate, when trying to find universal measures for all speakers; however, the result, in itself, is interesting.

Several possible explanations can account for the lack of significant variance in the measures applied to speaker g3. One possibility is that stress affects speakers in different ways, and some speakers may not show stress effects in their voice. A second possibility is that this speaker was not "stressed" by the experimental environment and so was not exhibiting stressful effects. A third possibility is that measures of stress are in the voice beyond the eight that we measured, and these additional measures are required for identifying stress in this speaker. Certainly, the last possibility deserves additional research. This study looked only at measures related to the excitation signal of speech, i.e., the low-frequency pitch pulse. Many higher frequency components of the voice may well reflect stress effects. These potential measures of stress deserve consideration. The other possibilities should be kept in mind when performing such research. Inducing stressful effects in an experimental setting is difficult, and the type of stress applied can vary. Also, different levels of stress may be required to induce the same degree of stressful effect in different speakers.

The following discussion addresses the within speaker analysis summary table by considering each measure individually.

Amplitude. The Amplitude column contains no "+" or "-" significance markers. In short, amplitude was not a significant measure for any speaker. This is consistent with our finding across speakers; that is, there was no significance. Amplitude was significant for distinguishing speakers, but not for distinguishing stress. This suggests that the loudness of these speakers' voices did not change significantly under this type of stress.

Opening Slope. The measure of opening slope varies significantly between stressed and normal speech for two speakers, b1 and n2. For both speakers, it lowers for stressed speech. The F-Value is relatively high for both speakers, suggesting that this measure is a reliable measure for these speakers.

Closing Slope. The closing slope varied significantly for six speakers (b1, b3, g2, n1, n2, and n3). This was the second most widespread measure; that is, it is second only to BB as the measure that proved significant for the most number of speakers. This correlates with the earlier finding across speakers, that closing slope was significant. The significance of this measure varies by speaker, with F-Values ranging from a low of 5 for speaker n1 to 28 for speaker g2. An F-Value of 4 is just barely significant, corresponding to a P-Value of approximately 0.045; whereas, an F-Value of 12 or higher corresponds to a P-Value of approximately 0.001 or higher. Again, any P-Value below 0.05 is considered significant at the 95-percent confidence level.

Ratio. The ratio measure was found significant for five speakers (b2, b3, g1, n2, and n3) in the summary table. The ratio lowered for two speakers (b2 and b3) under stress and rose for the other three speakers. The F-Values were not high for four of the speakers, suggesting ratio was only strongly significant for one speaker, n2. Overall, we conclude that ratio is significant for a majority of speakers, but not highly reliable due to the low F-Values.

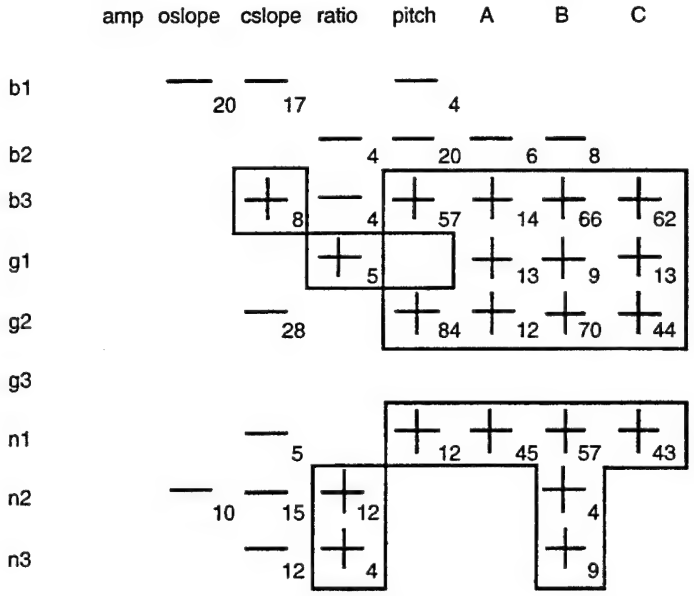
Pitch. The summary table shows pitch to be a measure whose mean value varied significantly between stressed and normal speech for five speakers (b1, b2, b3, g2, and n1). For three of these speakers, pitch rose under stress; and for two of the speakers, pitch lowered. Pitch appears to be a fairly reliable measure, since F-Values exceeded 10 for 4 of the 5 speakers (and exceeded 50 for 2 of the speakers). The earlier across-speaker analysis also found pitch to be a significant measure. In short, pitch appears to be a reliable measure for distinguishing stressed speech in the majority of speakers; however, the pitch may rise or lower under stress, and which occurs is speaker-dependent.

AA. AA was a significant parameter for five speakers (b2, b3, g1, g2, and n1), rising for four of the speakers, lowering only for speaker b2. The F-Values were all relatively high, with the lowest at 6, and the others above 10. AA reflects the amplitude adjustment to the pulse, which needs to rise when BB and CC rise in order to maintain a normalized pulse height. Thus, AA is related to BB and CC. All three parameters, AA, BB, and CC were found significant in the across speaker analysis. In short, the summary table suggests AA is a reliable measure for distinguishing stress in a majority of the speakers studied.

BB. The mean BB value varies significantly between stressed and normal speech for seven speakers (b2, b3, g1, g2, n1, n2, and n3). This is the most widespread significance of any of the measures, i.e., it was significant for most of the speakers. The F-Values, above 50 for 3 speakers, were marginal for only one speaker, n2, at 4. BB, which was found earlier to be a significant measure across speakers, corresponds with the front rising portion of the beta function. BB would change, if the front half of the pulse changed in some global sense, rather than in a local ripple or curve in some portion of the pulse. In general, BB appears to be a reliable measure for the majority of speakers.

CC. The mean of the final measure, CC, varied significantly between stressed and normal speech for four of the nine speakers, rising in all four cases. Three of the 4 F-Values were above 40, and the one lower was still at 13. CC, which was found to be a significant parameter in the across speaker analysis, corresponds with the second, closing portion of the beta function. It would change in response to any global change in this closing portion, as opposed to a local ripple or curve. In short, CC was a significant parameter for almost half the speakers studied; and the F-Values suggest it is quite reliable.

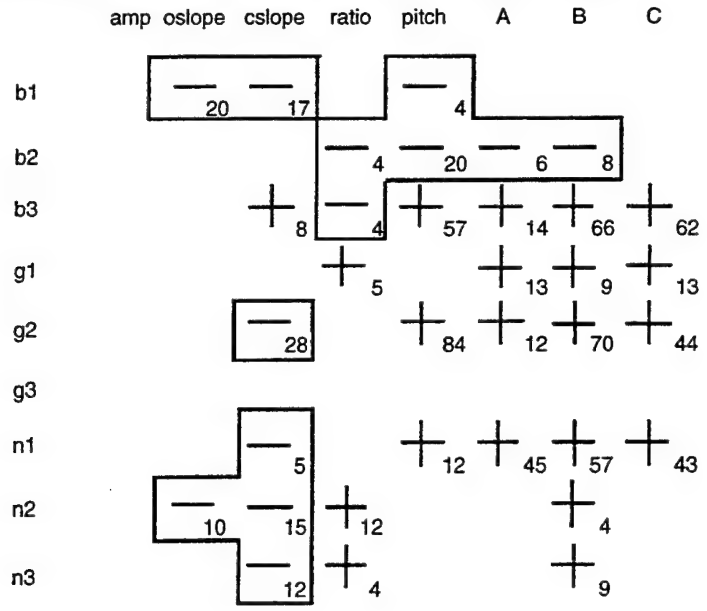
Within Speaker Summary Table: General Observations. Certain trends in the results described in the summary table can be seen, if we look at "+" measures versus "-" measures. In the following summary table, the "+" values are isolated:



Before proceeding, note that the shape of the "+" arrangements across adjacent rows is not significant, since the rows have no particular order. The first significant observation is that, for a number of speakers, AA, BB, and CC tend to change together and change in the same positive direction. When beta functions are plotted with AA, BB, and CC values modified, one can easily see the following: when all three rise, the net effect is that the beta function "narrows." This suggests that the glottal pulses from stressed speech tend to narrow for many speakers. This narrowing might result from tension in the vocal folds, which causes them to open and close faster than normal.

Another interesting observation is that BB tends to rise and become significant, even when AA and CC are not. Again, adjusting beta functions shows that when BB rises, without CC rising, the pulse tends to shift to the right, resulting in a slower rising slope and a steeper closing slope. This suggests that the vocal folds are opening more slowly and/or closing more rapidly when a person is under stress.

The following shows the summary table with the "-" marks isolated:



Here, the closing slope tends to lower for most speakers, suggesting, as with the AA, BB, and CC parameters, that the closing slope is steeper either in response to a pulse narrowing or shifting to the right. Also note that for one speaker, b1, opening slope and closing slope changed, but AA, BB, and CC did not. Finally, AA, BB, and CC appear to capture large pulse changes, while the opening and closing slope measure local changes in slope at the rising and closing mid-points.

CHAPTER 9
NEURAL NETWORK EXPLORATION

A neural network is a nonlinear processing structure used widely for classification, system modeling, and detection. The structure of neural networks was inspired by scientists who wanted to model the human brain. Neural networks match the functionality of the brain in a very fundamental manner.

A neuron, seen in figure 5, is the fundamental element in a nervous system; and in particular, the brain. Each neuron in the brain receives and combines signals from many other neurons and produces an output (either firing or not firing), based on a weighted combination of these inputs. If the combined input is strong enough, the neuron fires; if not, it doesn't. Firing versus not firing can be considered analogous to the binary 0-1 situation that computers are based on.

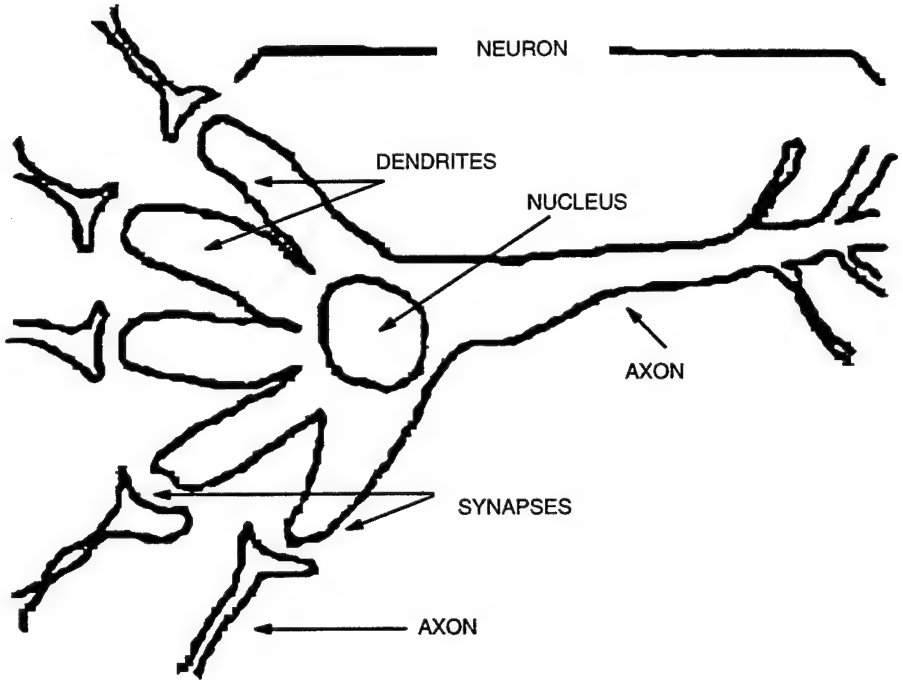


Figure 5. Physical structure of neuron.

Brains consist of billions of neurons interconnected by synapses (chemical conduits) from outputs (axons) to inputs (dendrites). These connections, along with the processing of the neuron, form the basic memory mechanism of the brain.

The artificial neuron analogy to the physiological system (figure 5) can be seen in figure 6, which shows a neural network processing element (or PE). Here the input is the vector **X**=(x1, x2, x3 ... xn), which is connected through the weight vector **W**=(w1, w2, w3 ... wn) to the neuron, which produces an output, **Y**. The neuron performs a thresholding on the weighted sum of the input vectors and produces a 0 or 1 output. A neural network consists of many neurons connected together in layers, as in figure 7.

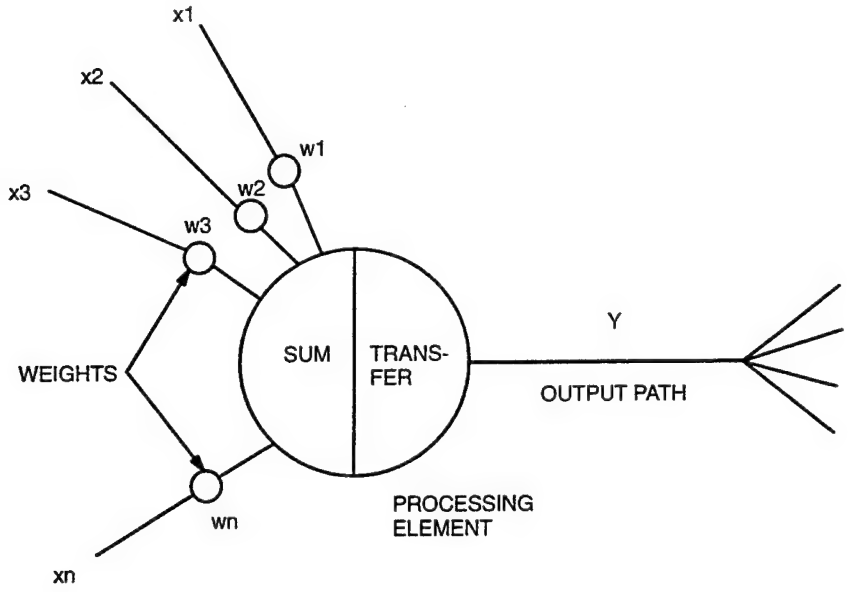


Figure 6. Neural network processing element.

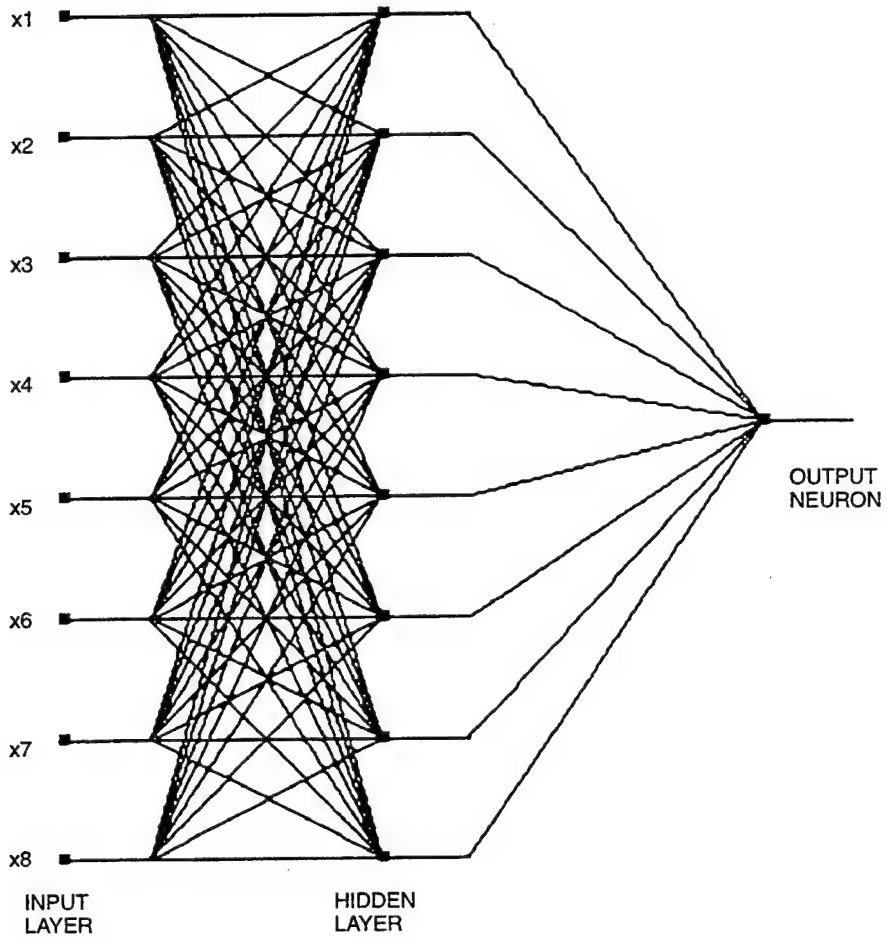


Figure 7. Neural network structure used for glottal pulse classification.

Network operation can be broken up into two phases: training and testing. Training consists of trying to find the optimal weight matrix (i.e., the set of weight values that best classifies the data of interest). The network is trained by presenting a set of feature vectors (X), calculating the output, determining the error, and readjusting the weight matrix. This process is repeated until the error is below some desired threshold. Several algorithms exist for readjusting the weight matrix. We used a very popular weight-update algorithm for classification paradigms, known as back propagation.

Once the network is trained, and the desired weight matrix is obtained, the network can be tested on new data to see how well it performs. Testing the network consists of presenting a set of feature vectors (**X**), calculating the output, and determining the cumulative error as the summation of the square of the actual output, minus the desired output.

The structure in figure 7 is the same as the neural network used for classifying the glottal pulses. It has eight PEs in the input layer, eight PEs in the hidden layer, and one PE in the output layer. Heuristics and pilot experiments were used to determine that this structure gave the best results. The back propagation algorithm demands a monotonically increasing transfer function, so the error can be propagated back through the network. Common transfer functions are hyperbolic tangent and log sigmoidal. We used the latter. The ouput of the last PE is hard limited to 0 or 1, mapping < 0.5 to 0 and > 0.5 to 1, to make a classification decision. On the output neuron, a 0 was considered normal, and a 1 was considered "stressed."

The feature vectors described earlier were used as the input vectors, and all experiments were conducted across all speakers. All eight features were used, with half the data saved for testing and half used for training. All the data were normalized by feature, so that the input values were between 0-1. The network was trained and tested several times. Each time the test and train data were picked randomly.

Over 25 trials, the mean classification rate was 71 percent, with a standard deviation of 2.2 percent. If guessing, we would expect to get 50-percent correct. Therefore, this is obviously better than guessing and indicates some information is being retrieved that indicates a level of stress.

CHAPTER 10
CONCLUSIONS

The major goals of the research were satisfied. A semiautomatic technique, capable of full automation, was developed for (1) extracting glottal pulses from speech, (2) inverse filtering to extract the glottal pulse, and (3) measurement of the eight pulse parameters. The statistical analysis confirmed that excitation-pulse parameters, closing slope, pitch, AA, BB, and CC are significant indicators of stress across speakers. The analysis suggested that pulses are narrowed or shifted to the right for most speakers under stress; and that pitch, ratio, and closing slope are significant parameters within speaker. Nonetheless, measures vary by speaker in important ways, and only amplitude appeared as an insignificant measure for all speakers.

One of the major findings was that, although some pulse parameters are significant across speakers, most of the parameters vary in speaker-specific ways. This suggests that stress-identification systems will need to use speaker-specific baselines for measuring variations.

We also found that neural-processing techniques could be used to achieve significant stress detection. Our simple neural-processing techniques were applied across speakers, without even considering speaker-specific findings, and achieved 70-percent stress identification (on a 50-percent, yes/no classification test).

These findings suggest two types of follow-on efforts. The first is a research effort to confirm whether other potential stress measures, in the spectral component of speech, are reliable indicators of stress. The second recommended effort is a prototype development of a stress-detection system, using the automatic glottal-pulse extraction, with baseline measures of normal speech for target speakers, and variation thresholds on parameters deemed significant from this study.

CHAPTER 11
REFERENCES

Alku, Paavo. 1992. "An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform," *Proceedings IEEE ICASSP* (pp. 29–32).

Cummings, K. E. and Clements, M. A. 1990. "Analysis of Glottal Waveforms Across Stress Styles," *Proceedings of the IEEE ICASSP* (pp. 369–372).

Cummings, K. E. and Clements, M. A. 1992. "Improvements to and Applications of Analysis of Stressed Speech Using Glottal Waveforms," *Proceedings of the IEEE ICASSP* (pp. 25–28).

CHAPTER 12
BIBLIOGRAPHY

Brenner, M., H. H. Branscomb, and G. E. Schwartz. 1979. "Psychological Stress Evaluator—Two Tests of a Vocal Measure," *Psychophysiology*, v. 16, no. 4 (July) (pp. 351-357).

Fuller, B. F. 1984. "Reliability and Validity of an Interval Measure of Vocal Stress," *Psychological Medicine*, v. 14, no. 1 (Feb) (pp. 159-166).

Flanagan, J. L. 1972. *Speech Analysis, Synthesis and Perception*. Springer-Verlag (pp. 232-246).

Griffin, G. R. 1987. "The Effects of Different Levels of Task Complexity on Three Vocal Measures," *Aviation, Space and Environmental Medicine*, v. 58 (Dec) (pp. 1165-1170).

Gunn, J. and G. Gudjonsson. 1988. "Using the Psychological Stress Evaluator in Conditions of Extreme Stress," *Psychological Medicine*, v. 18, no. 1 (Feb) (pp. 235-238).

Kuroda, I., O. Fugiwara, N. Okamura, and N. Utsuki. 1976. "Method for Determining Pilot Stress Through Analysis of Voice Communication," *Aviation, Space, and Environmental Medicine*, v. 47, no. 5 (May) (pp. 528-533).

O'Shaughnessy, D. 1987. *Speech Communication: Human and Machine*. Addison-Wesley Publishing Company (pp. 50-55).

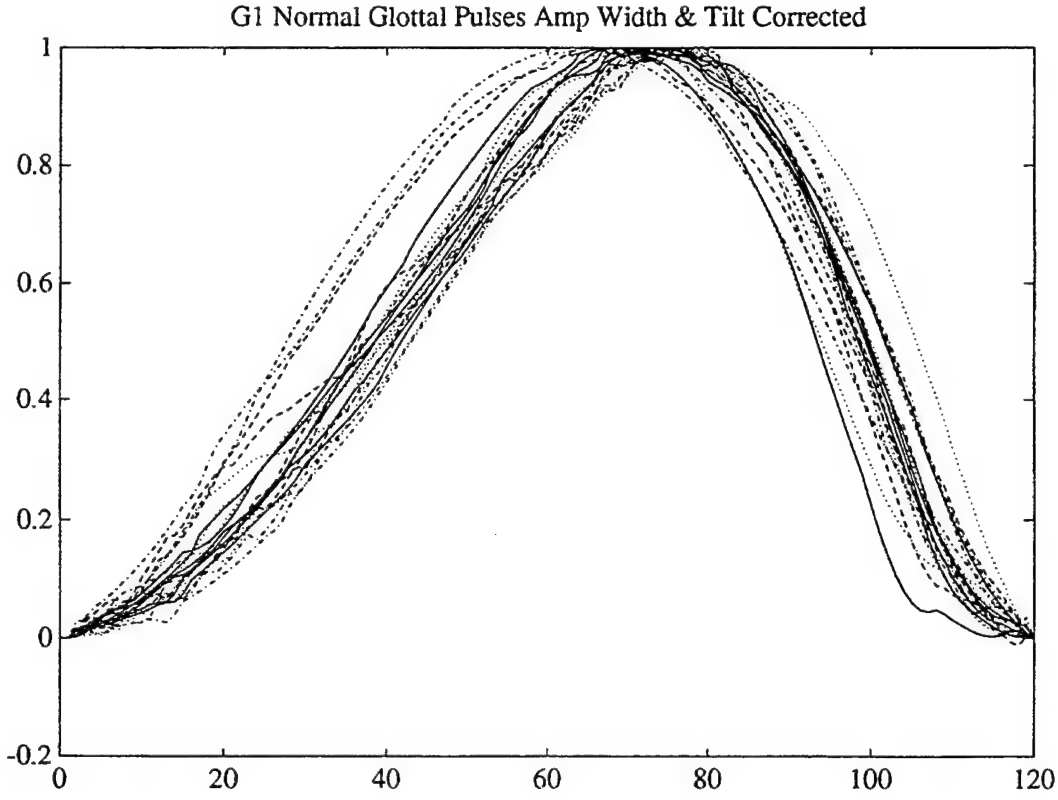
Rosenberg, A. E. 1971. "Effect of Global Pulse Shape on the Quality of Natural Vowels," *Journal of the Acoustical Society of America*, v. 49, no. 2 (part 2, 1971) (pp. 583-588).

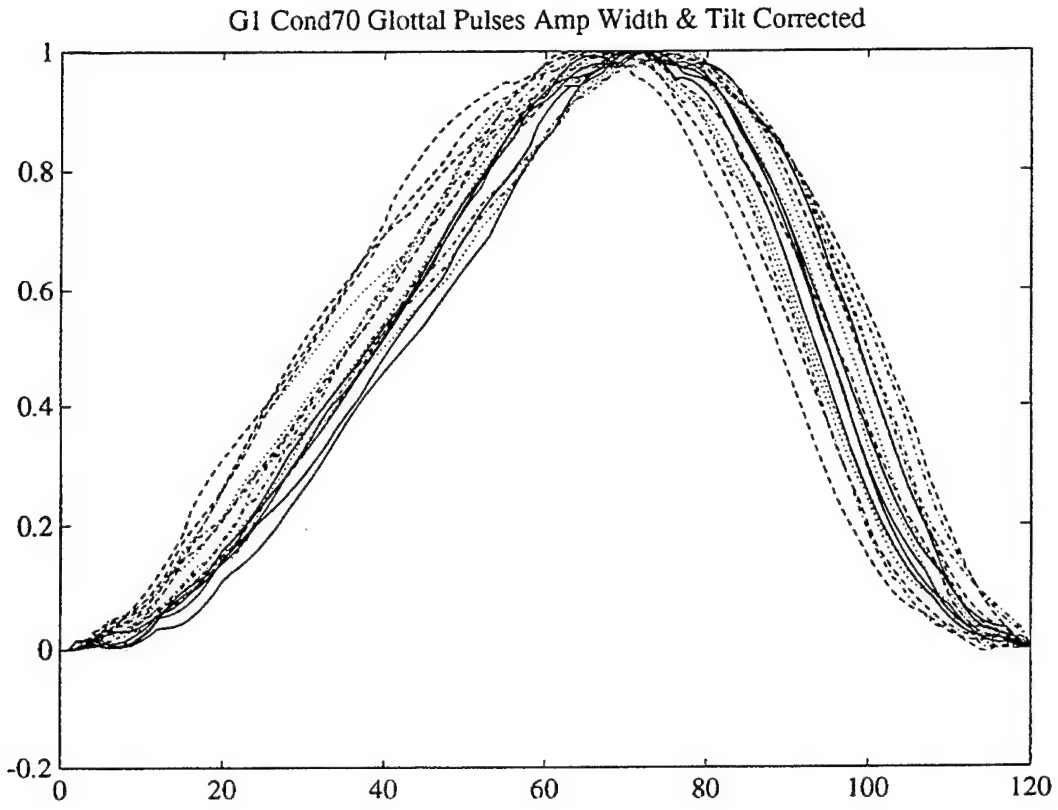
Tolkmitt, F. J., and K. R. Scherer. 1986. "Effect of Experimentally Induced Stress on Vocal Parameters," *Journal of Experimental Psychology: Human Perception and Performance*, v. 12, no. 3 (pp. 302-313).

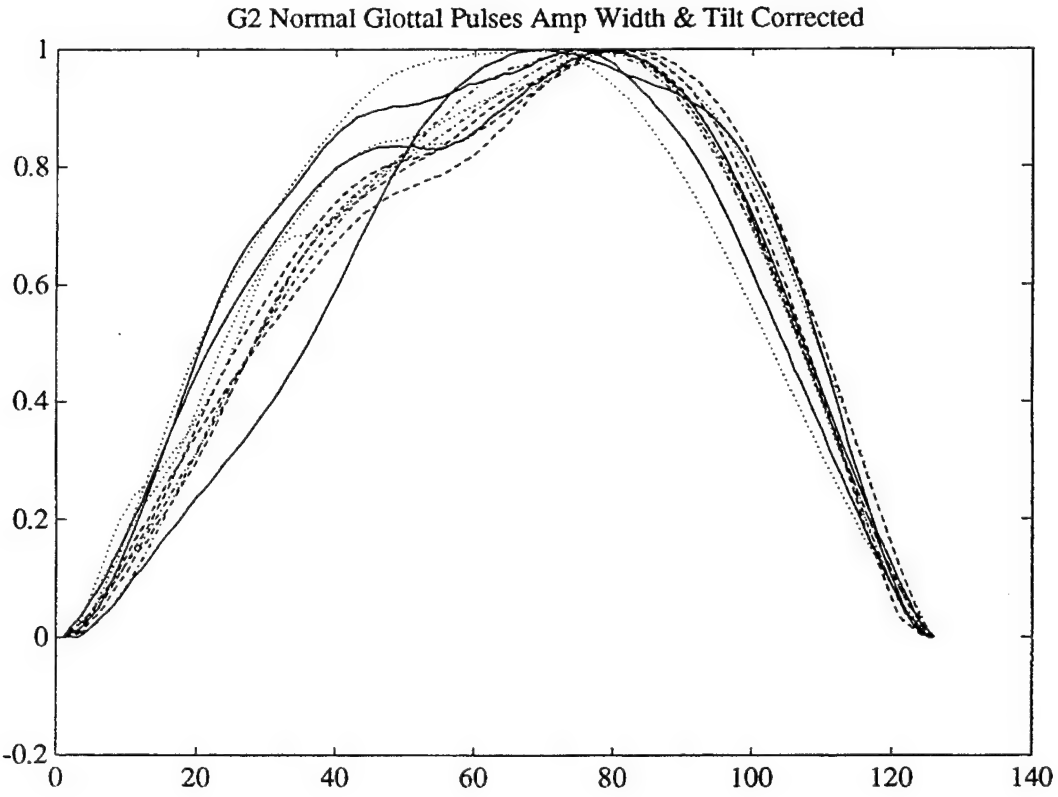
Wong, D. Y., J. D. Markel, and A. H. Gray, Jr. 1979. "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP–27, no. 4 (Aug) (pp. 350-355).

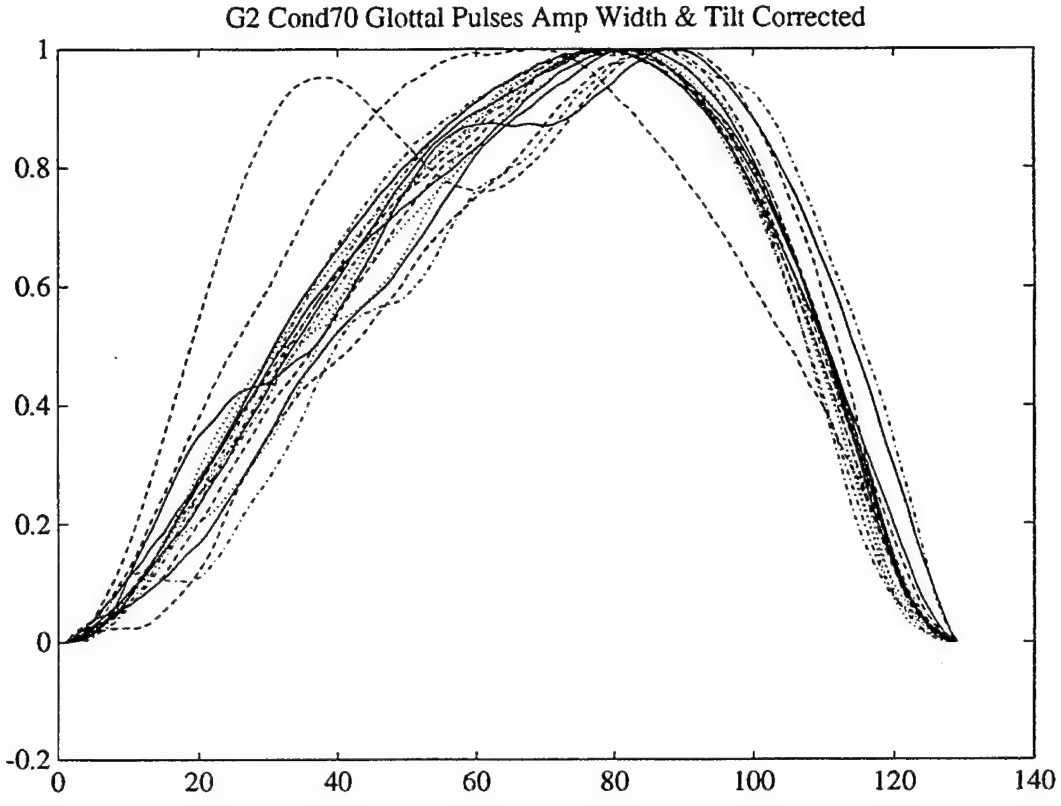
APPENDIX A
THE GLOTTAL PULSES

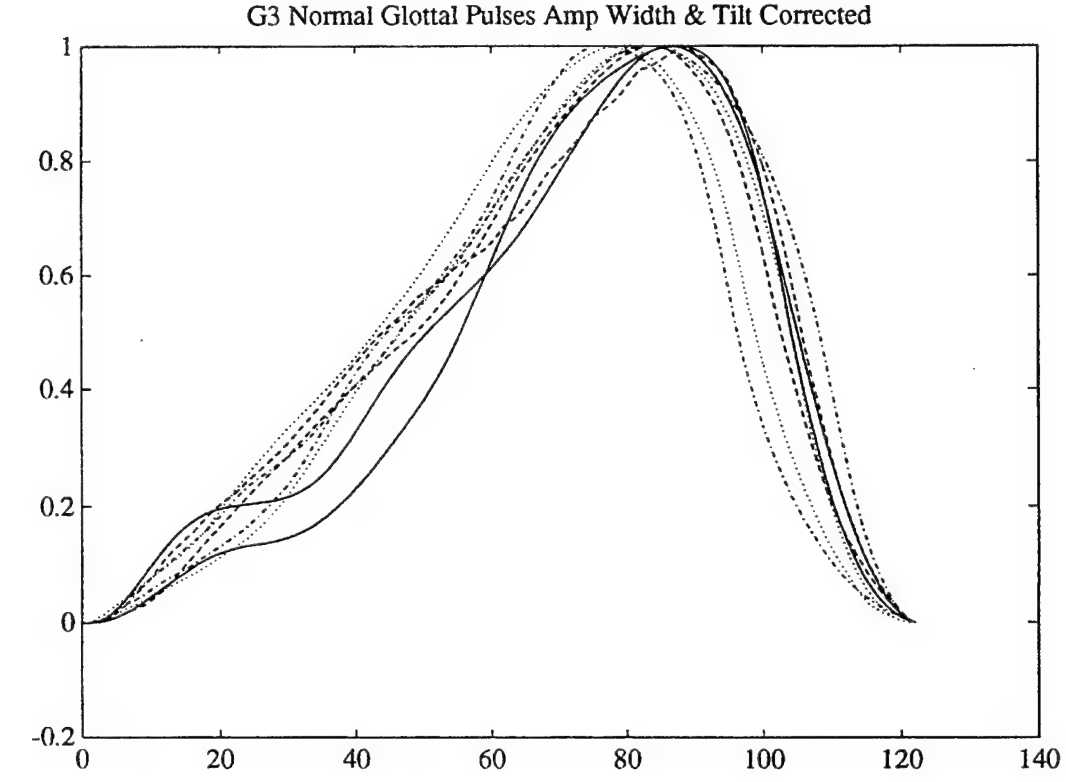
The following charts depict the glottal pulses used in this study. Nine speakers are included (g1, g2, g3, b1, b2, b3, n1, n2, and n3). There is a chart for each speaker under each of two speaking styles, stressed (cond70) and normal. The pulses are shown after normalization for amplitude, width, and tilt.

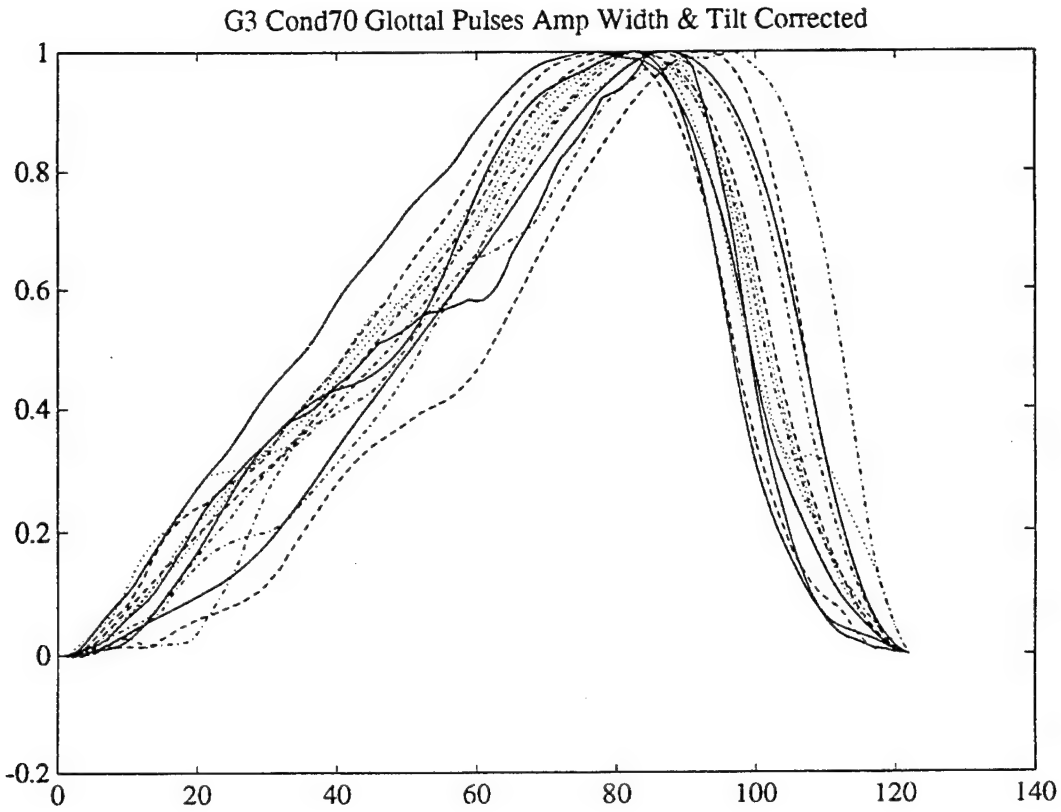


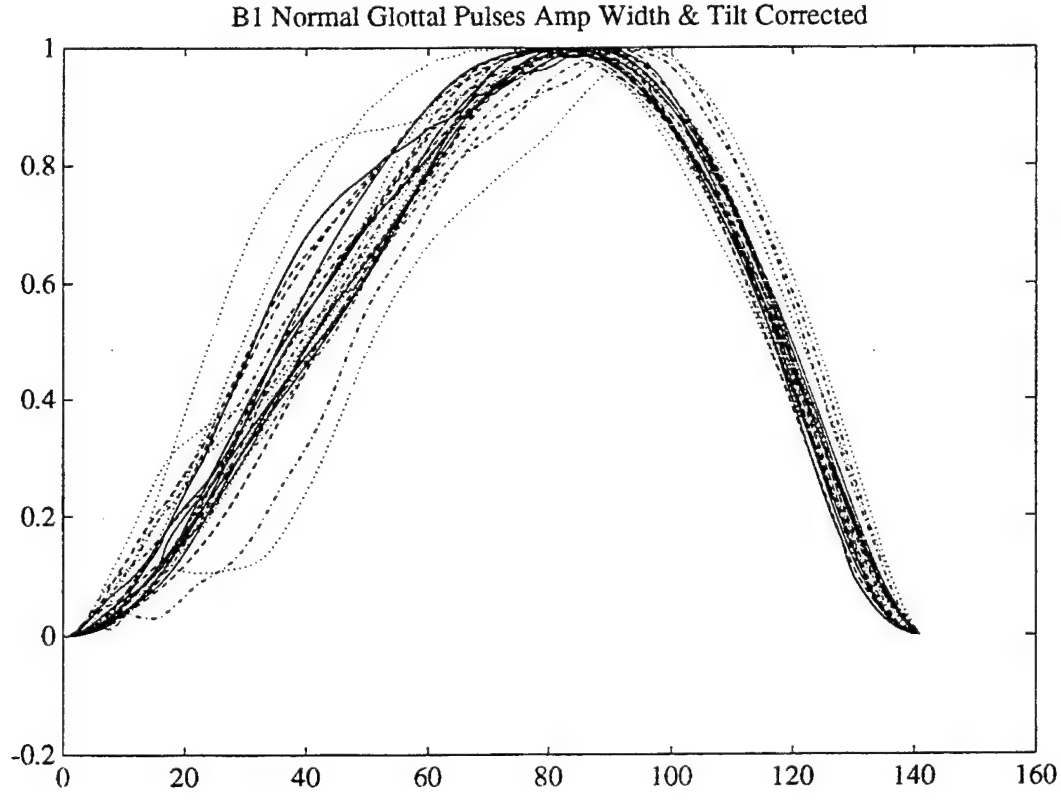

G1 Cond70 Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120


G2 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140

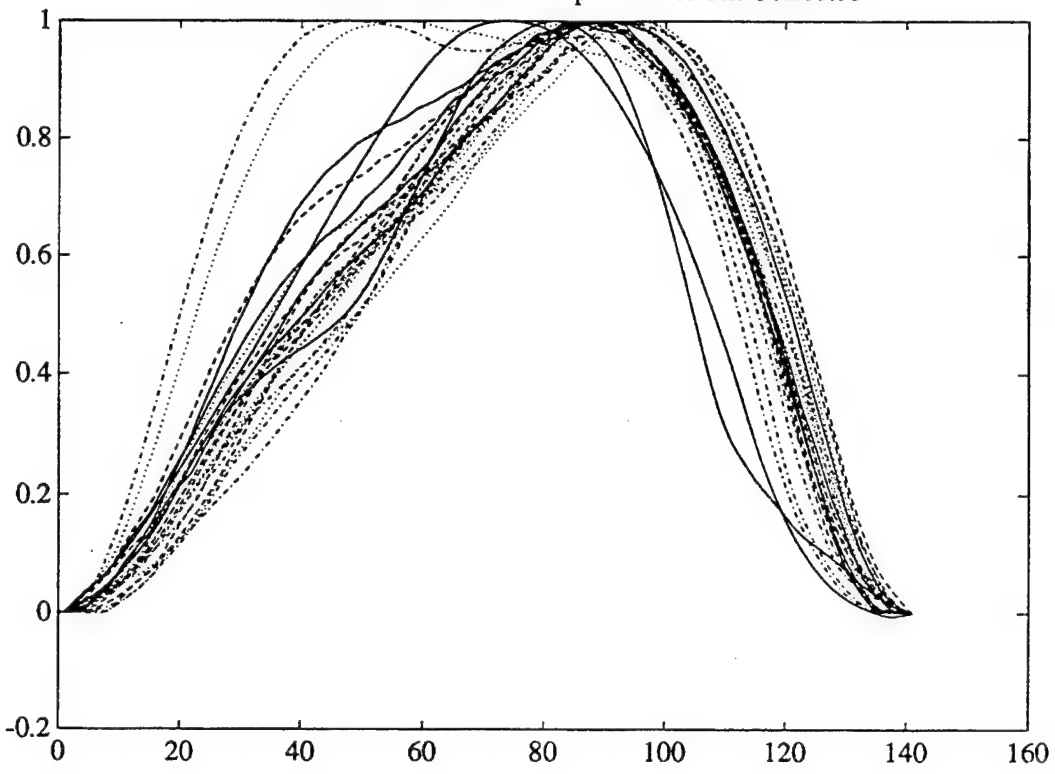

G2 Cond70 Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140


G3 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140

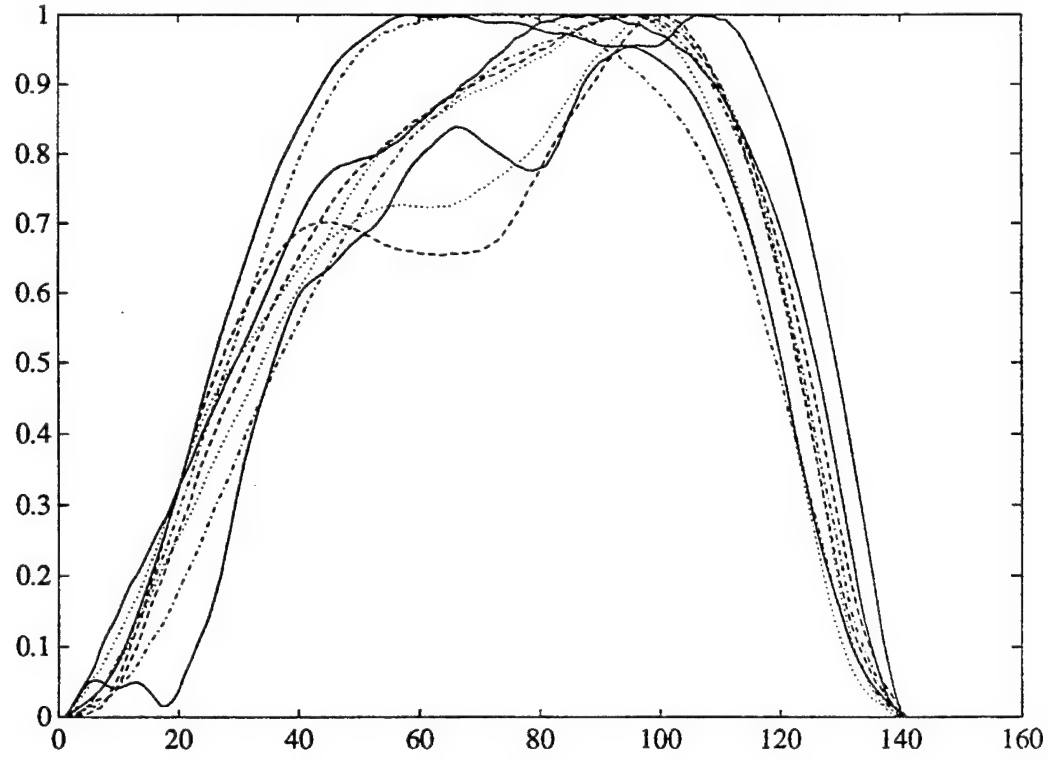

G3 Cond70 Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140

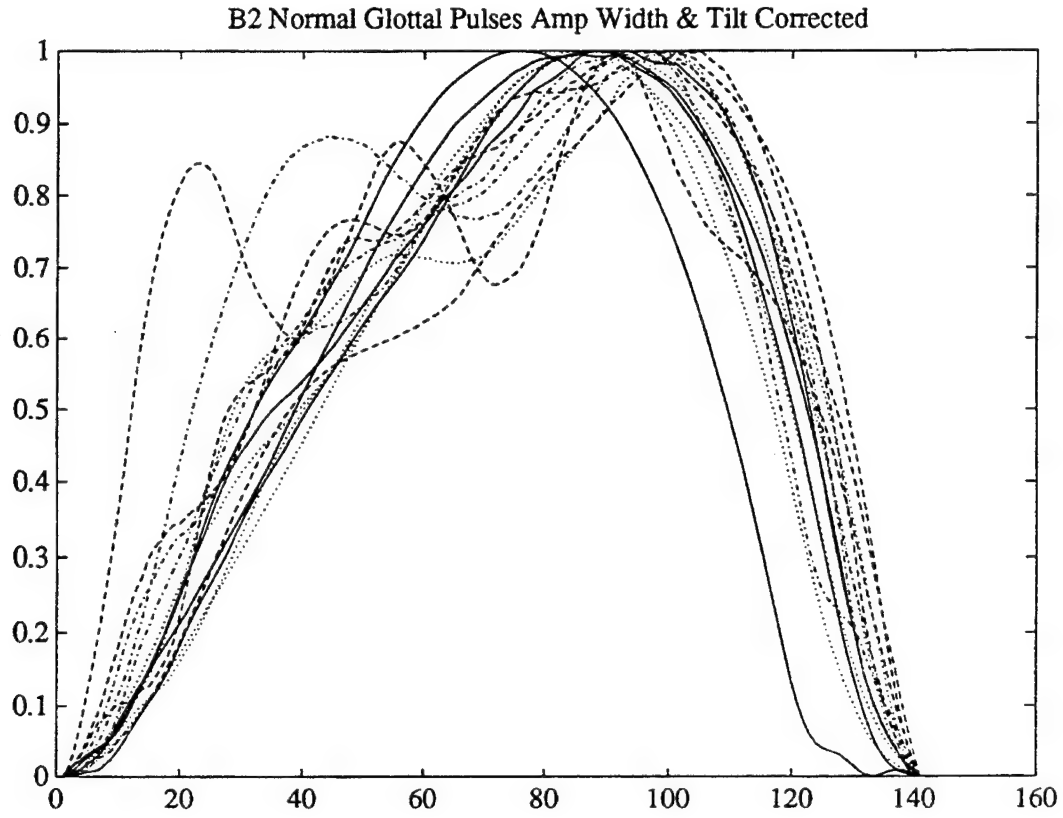

B1 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160

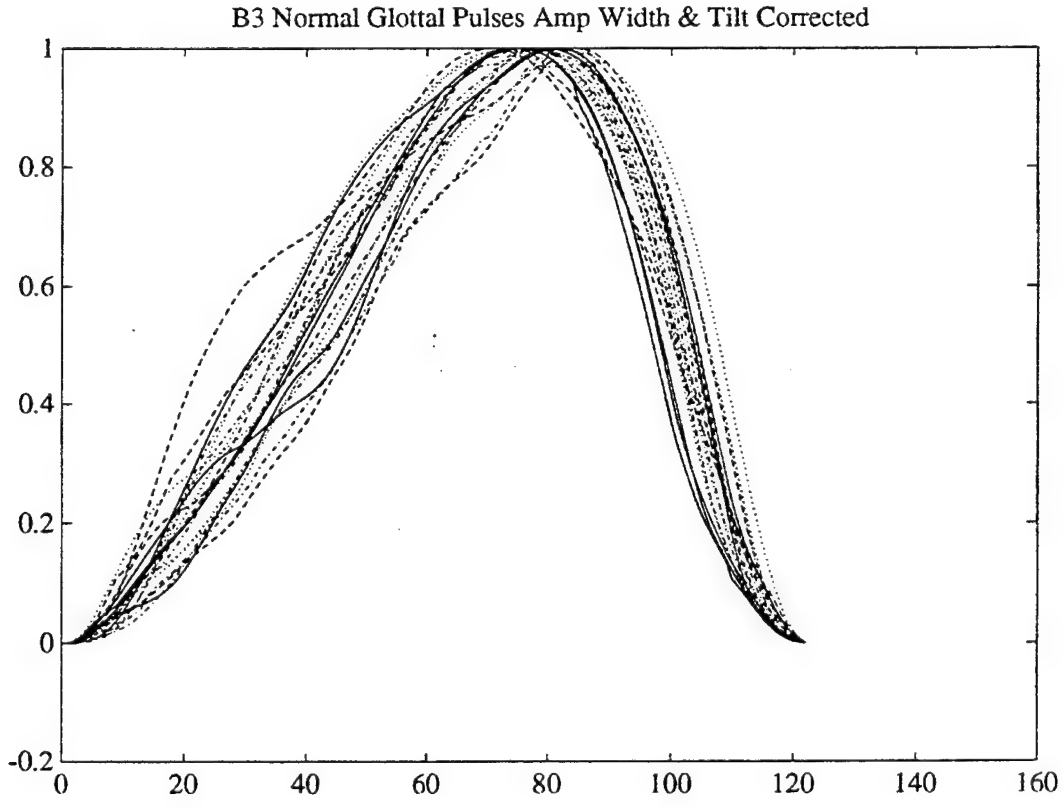
B1 Cond70 Glottal Pulses Amp Width & Tilt Corrected



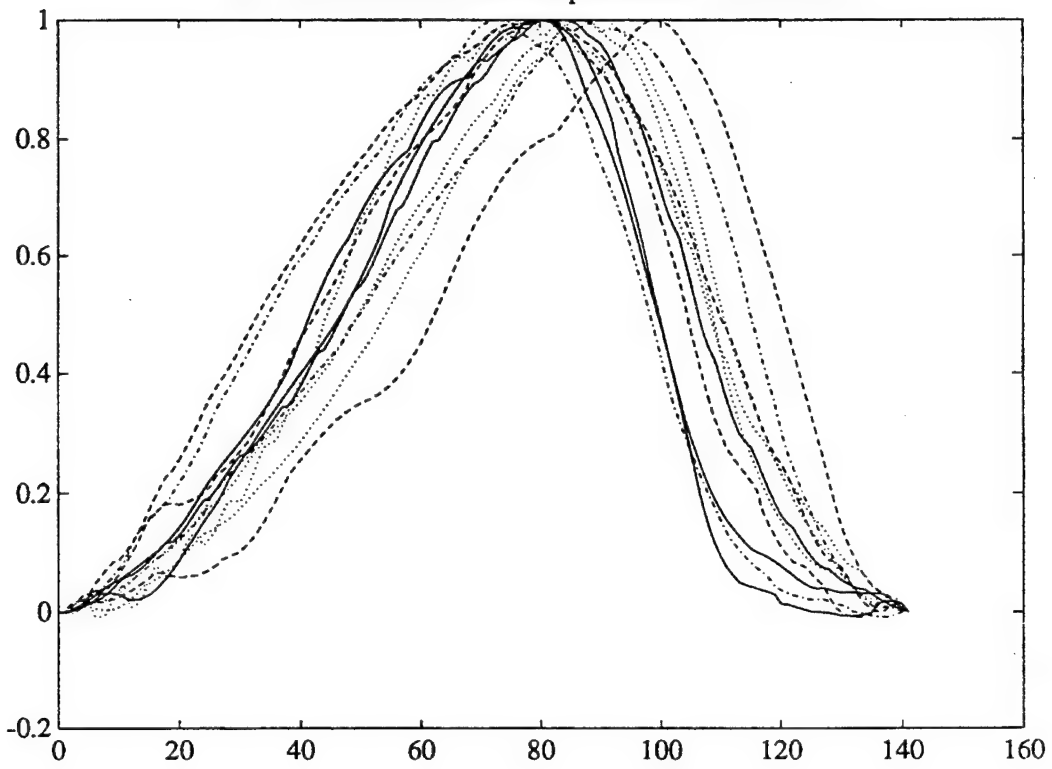
B2 Cond70 Glottal Pulses Amp Width & Tilt Corrected



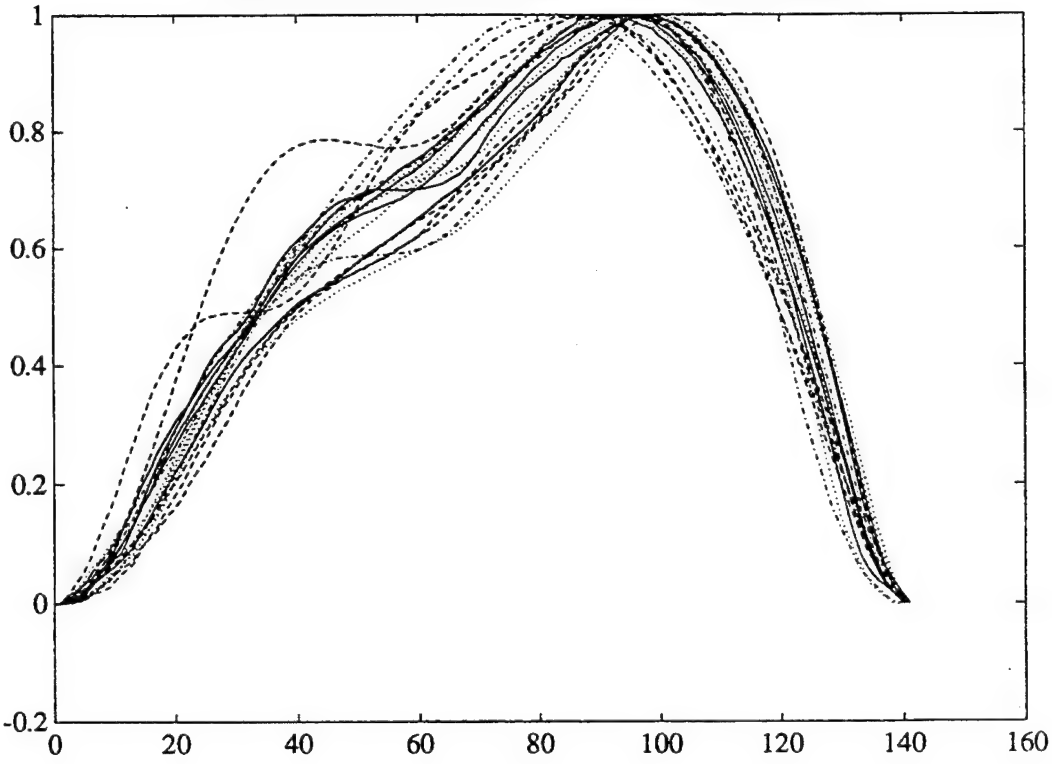

B2 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160

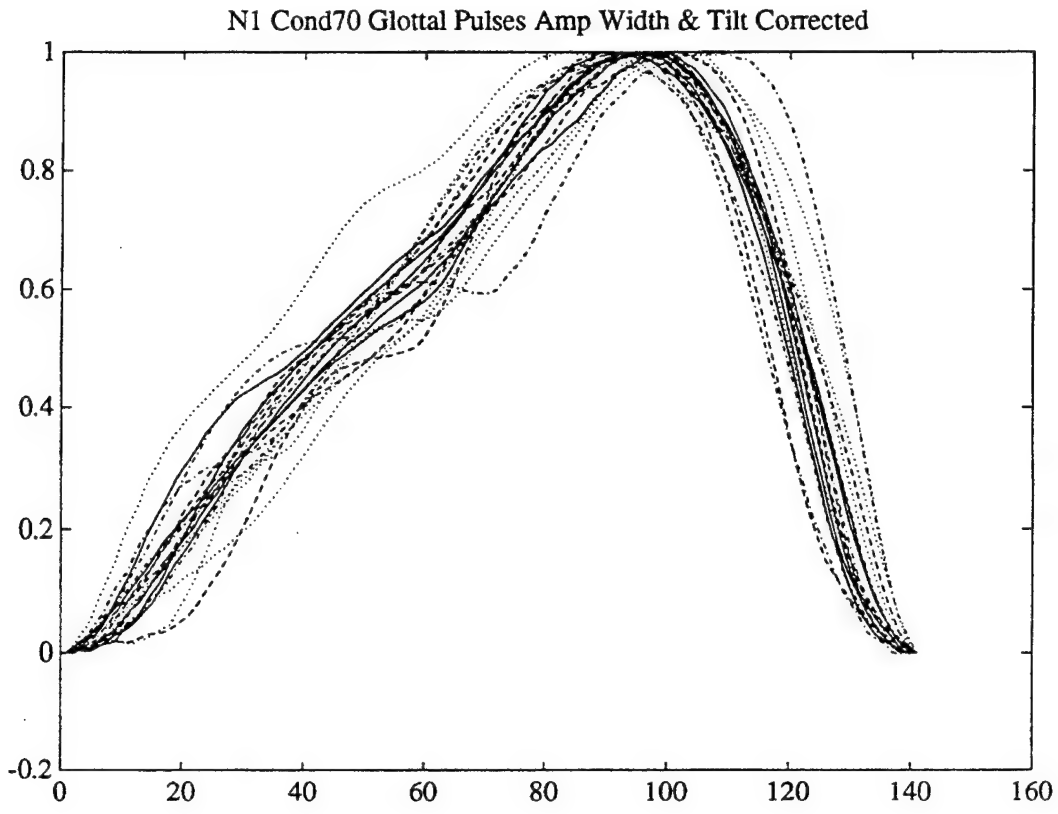

B3 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160

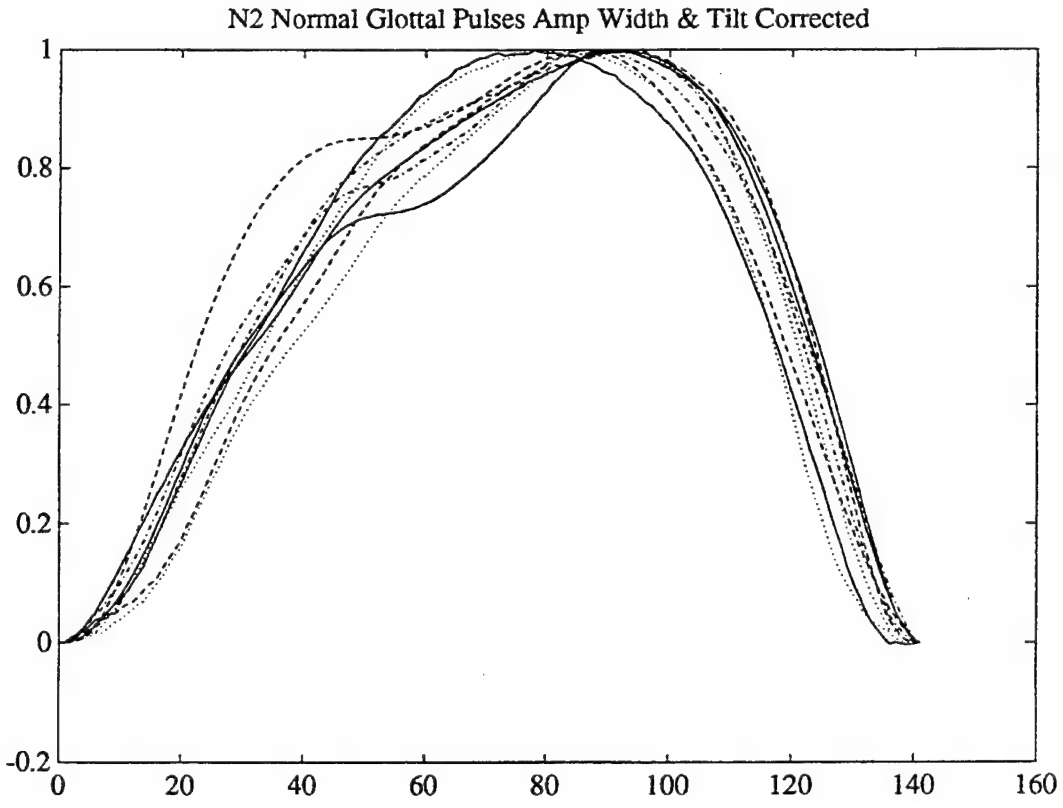
B3 Cond70 Glottal Pulses Amp Width & Tilt Corrected

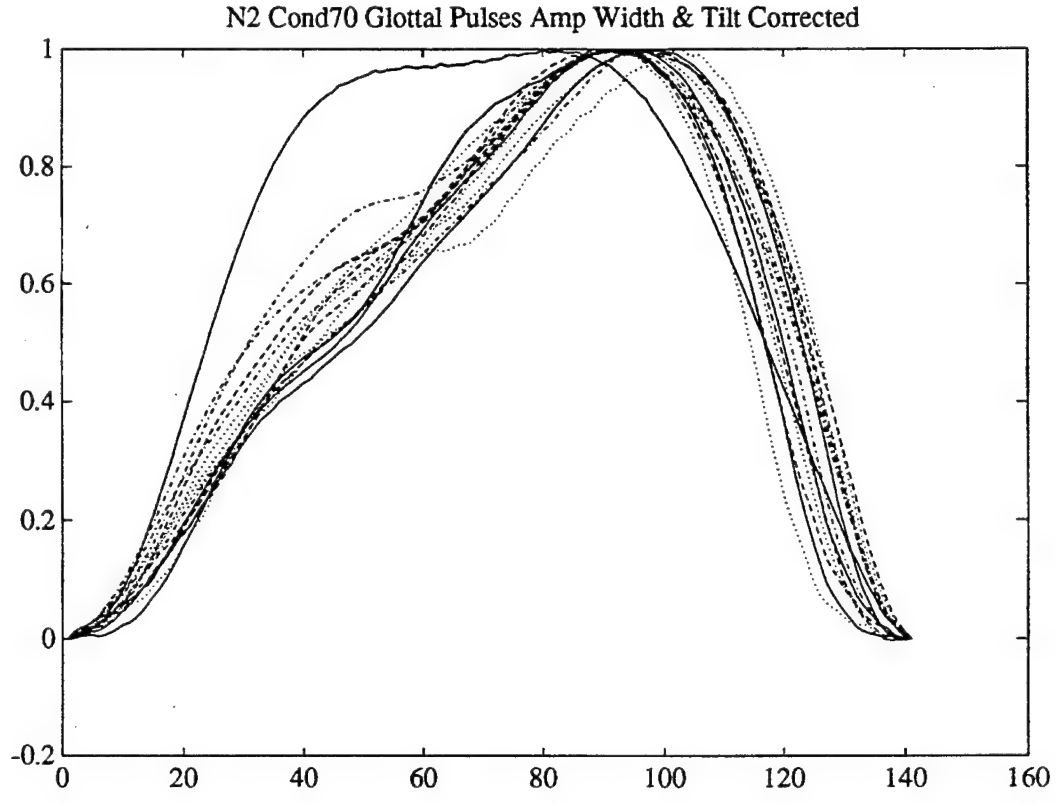


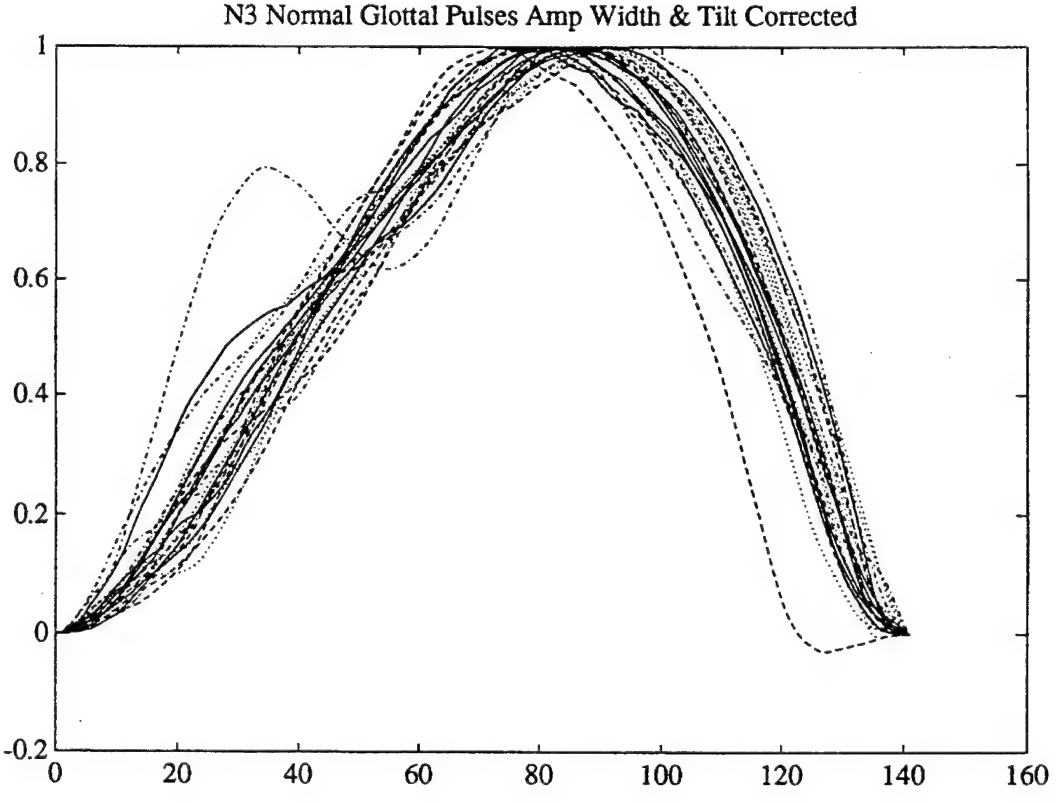
N1 Normal Glottal Pulses Amp Width & Tilt Corrected

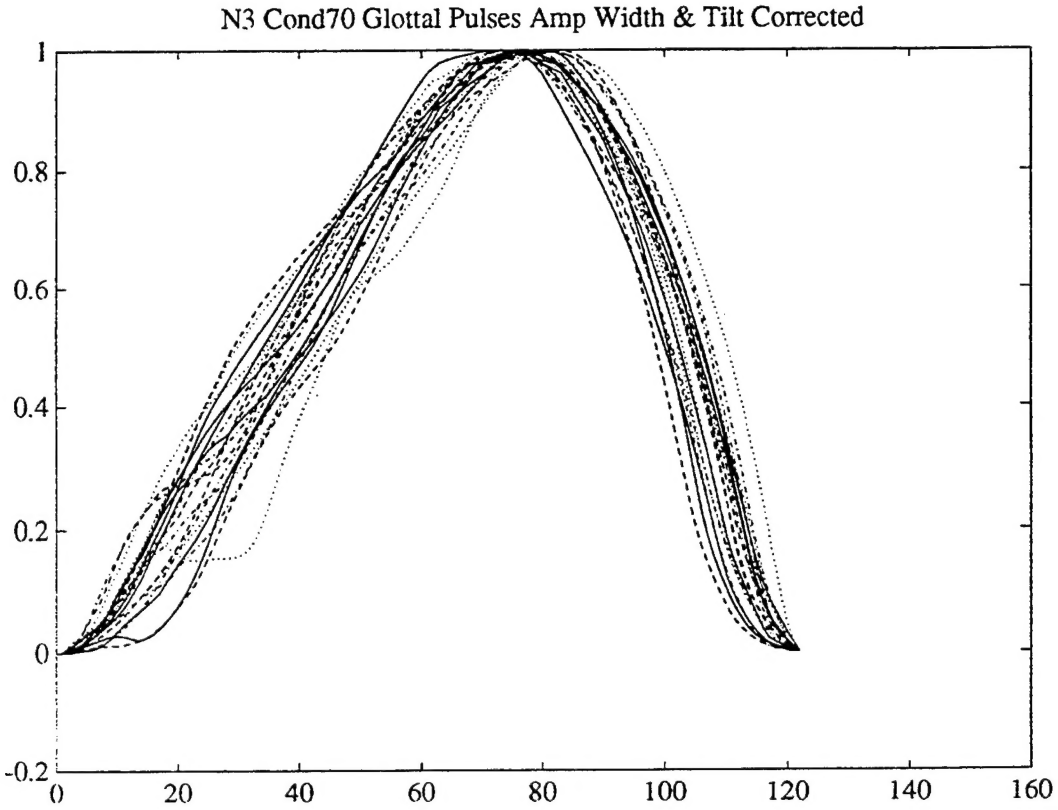



N1 Cond70 Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160


N2 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160


N2 Cond70 Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160


N3 Normal Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160


N3 Cond70 Glottal Pulses Amp Width & Tilt Corrected
1
0.8
0.6
0.4
0.2
0
-0.2
0
20
40
60
80
100
120
140
160

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1994	Final: FY 93 – FY 93

4. TITLE AND SUBTITLE
DETECTION OF STRESS BY VOICE: ANALYSIS OF THE GLOTTAL PULSE

5. FUNDING NUMBERS
PE: 0602936N
PR: ZF2301
SUBPROJ: RV36I21
WU: DN303026

6. AUTHOR(S)
Jeff Waters, Steve Nunn, Brenda Gillcrist, Eric VonColln

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Command, Control and Ocean Surveillance Center (NCCOSC)
RDT&E Division
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER
TR 1652

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of the Chief of Naval Research
OCNR–20T
Arlington, VA 22217–5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

An Independent Exploratory Development (IED) study was performed to determine whether or not significant measures are present in the human voice for detecting the emotional reaction, "stress." A technique was implemented for automatically measuring parameters of the glottal pulse, to see which might be indicators of stress. The results of this IED study confirmed that several of the measures are significant indicators of stress; for example, the glottal pulse generally narrows or shifts in mass under stress. And although pitch and beta-function parameters were significant across speakers, the measures were largely speaker dependent.

14. SUBJECT TERMS
Mission Area: Command and Control

speech
voice processing
stress detection
glottal pulse

15. NUMBER OF PAGES
81

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

21a. NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE *(include Area Code)*	21c. OFFICE SYMBOL
Jeff Waters	(619) 553–3657	Code 44213

INITIAL DISTRIBUTION

Code 0012	Patent Counsel	(1)
Code 0271	Archive/Stock	(6)
Code 0274	Library	(2)
Code 40	R. C. Kolb	(1)
Code 44	J. D. Grossman	(1)
Code 44213	J. H. Waters	(30)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (4)

NCCOSC Washington Liaison Office
Washington, DC 20363–5100

Navy Acquisition, Research and Development
Information Center (NARDIC)
Arlington, VA 22244–5114

GIDEP Operations Center
Corona, CA 91718–8000